**Functional and evolutionary characterization of flowering-related long non-coding RNAs**

DISSERTATION

zur Erlangung des akademischen Grades

Doctor rerum naturalium

(Dr. rer. nat.)

im Fach Biologie

eingereicht an der

Lebenswissenschaftlichen Fakultät

der Humboldt-Universität zu Berlin

Von

M.Sc. Li Chen

Präsidentin der Humboldt-Universität zu Berlin

Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät

Prof. Dr. Bernhard Grimm

Gutachter/innen:

Prof. Dr. Kerstin Kaufmann

Prof. Dr. Susann Wicke

Prof. Dr. Dorothee Staiger

Tag der mündlichen Prüfung: 23.04.2021

# Table of contents

**Zusammenfassung**

Genomweite Bemühungen haben eine große Anzahl langer nichtkodierender RNAs (lncRNAs) identifiziert, obwohl ihre möglichen Funktionen weitgehend rätselhaft bleiben. Hier verwendeten wir ein System zur synchronisierten Blüteninduktion in *Arabidopsis*, um 4106 blütenbezogene lange intergene RNAs (lincRNAs) zu identifizieren. Blütenbezogene lincRNAs sind typischerweise mit funktionellen Enhancern assoziiert, die bidirektional transkribiert werden und mit verschiedenen funktionellen Genmodulen assoziiert sind, die mit der Entwicklung von Blütenorganen zusammenhängen, die durch Koexpressionsnetzwerkanalyse aufgedeckt wurden. Die Master-regulatorischen Transkriptionsfaktoren (TFs) APETALA1 (AP1) und SEPALLATA3 (SEP3) binden an lincRNA-assoziierte Enhancer. Die Bindung dieser TFs korreliert mit der Zunahme der lincRNA-Transkription und fördert möglicherweise die Zugänglichkeit von Chromatin an Enhancern, gefolgt von der Aktivierung einer Untergruppe von Zielgenen. Beispielsweise zeigt die lincRNA Aralnc.19175 (zuvor als *LINC-AP2* bezeichnet), die zwei durch AP1 und SEP3 gebundene Enhancer überspannt, eine zunehmende Aktivität, verbunden mit der zunehmenden Zugänglichkeit der beiden entsprechenden Enhancer. Wir nehmen an, dass die Enhancer-assoziierte lincRNA-Expression funktionell mit der TF-Aktivität bei der Regulation von Blütengenen zusammenhängt.

Darüber hinaus ist die Evolutionsdynamik von lincRNAs in Pflanzen, einschließlich nicht blühender Pflanzen, noch nicht bekannt, und das Expressionsmuster in verschiedenen Pflanzenarten war ziemlich unbekannt. Hier identifizierten wir Tausende von lincRNAs in 26 Pflanzenarten, einschließlich nicht blühender Pflanzen, und ermöglichten es uns, sequenzkonservierte und auf Syntenie basierende homologe lincRNAs abzuleiten und konservierte Eigenschaften von lincRNAs während der Pflanzenentwicklung zu untersuchen. LincRNAs in verschiedenen Pflanzen zeigen konservierte Eigenschaften und der Anteil von lincRNAs im Genom ist ungefähr linear proportional zur Genomgröße. Ein direkter Vergleich von lincRNAs zeigt, dass die meisten lincRNAs speziesspezifisch sind und das Expressionsmuster von lincRNAs einen hohen Transkriptionsumsatz nahe legt. Darüber hinaus zeigen konservierte lincRNAs eine aktive Regulation durch Transkriptionsfaktoren wie AP1 und SEP3. Konservierte lincRNAs zeigen eine konservierte blütenbezogene Funktionalität sowohl in der Brassicaceae- als auch in der Grasfamilie. Darüber hinaus sind TEs mit diesen konservierten lincRNAs assoziiert und fördern die Stabilisierung von lincRNAs während der Evolution von Pflanzen. Schließlich identifizierten wir auch konservierte lincRNAs in nicht blühenden Pflanzen und schlagen potenziell meristembezogene Funktionen vor. Die Evolutionslandschaft von lincRNAs in Pflanzen liefert wichtige Einblicke in die Erhaltung und Funktionalität von lincRNAs und stellt der Community Ressourcen für lincRNAs für weitere

experimentelle Untersuchungen zur Verfügung.

**Schlüsselwörter: lincRNAs, Blütenentwicklung, Genregulationsnetzwerk, Enhancer, TFs, Evolution, *Arabidopsis*, Reis, Mais.**

**Abstract**

Genome-wide efforts have identified a large number of long non-coding RNAs (lncRNAs), although their potential functions remain largely enigmatic. Here, we used a system for synchronized floral induction in *Arabidopsis* to identify 4106 flower-related long intergenic RNAs (lincRNAs). Flower-related lincRNAs are typically associated with functional enhancers which are bi-directionally transcribed and are associated with diverse functional gene modules related to floral organ development revealed by co-expression network analysis. The master regulatory transcription factors (TFs) APETALA1 (AP1) and SEPALLATA3 (SEP3) bind to lincRNA-associated enhancers. The binding of these TFs is correlated with the increase in lincRNA transcription and potentially promotes chromatin accessibility at enhancers, followed by activation of a subset of target genes. For example, the lincRNA *Aralnc.19175* (designated previously as *LINC-AP2*) spanning two enhancers bound by AP1 and SEP3 shows increasing activity, coupled with the increasing accessibility of the two corresponding enhancers. We hypothesize that enhancer-associated lincRNA expression is functionally linked with TF activity in floral gene regulation.

Furthermore, the evolutionary dynamics of lincRNAs in plants including non-flowering plants still remain to be elusive and the expression pattern in different plant species was quite unknown. Here, we identified thousands of lincRNAs in 26 plant species including non-flowering plants, and allow us to infer sequence conserved and synteny based homolog lincRNAs, and explore conserved characteristics of lincRNAs during plants evolution. LincRNAs in diverse plants demonstrate conserved characteristics and the proportion of lincRNAs in the genomes is roughly in linearly proportion with the genome size. Direct comparison of lincRNAs reveals most lincRNAs are species-specific and the expression pattern of lincRNAs suggests their high evolutionary gain and loss. Moreover, conserved lincRNAs show active regulation by transcriptional factors such as AP1 and SEP3. Conserved lincRNAs demonstrate conserved flower related functionality in both the *Brassicaceae* and grass family. Furthermore, TEs are associated with these conserved lincRNAs and drive stabilization of lincRNAs during the evolution of plants. Finally, we also identified conserved lincRNAs in non-flowering plants and suggests potentially meristem related functions. The evolutionary landscape of lincRNAs in plants provide important insights into the conservation and functionality of lincRNAs and provide lincRNAs resources with the community for further experimental investigation.

**Keywords: lincRNAs, flower development, gene regulatory network, enhancers, TFs, evolution, *Arabidopsis*, rice, maize.**

**Abbreviation**

| | |
|---|---|
| LincRNAs | Long intergenic non-coding RNAs |
| LncRNAs | Long non-coding RNAs |
| TF | Transcription Factor |
| TFBS | TF binding sites |
| RdDM | RNA directed DNA methylation |
| *LAIR* | *LRK Antisense Intergenic RNA* |
| eRNAs | Enhancer-associated lncRNAs |
| RNA pol II | RNA polymerase II |
| TEs | Transposable elements |
| TE-lncRNAs | Transposable element-associated lncRNAs |
| Non-TE-lincRNAs | Non-Transposable element-associated lncRNAs |
| ORFs | Open reading frames |
| SE | SERRATE |
| CBP20 | CAP BINDING PROTEIN20 |
| CBP80 | CAP BINDING PROTEIN80 |
| CBF1 | C-REPEAT/DRE BINDING FACTOR 1 |
| IDN2 | INVOLVED IN DE NOVO 2 |
| DRM2 | DORMANCY ASSOCIATED GENE 2 |
| SEP3 | SEPALLATA3 |
| HNRNPK | HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN K |
| U1 snRNP | U1 small nuclear ribonucleoprotein particle |
| CUTs | Cryptic unstable transcripts |
| PROMPTs | Promoter upstream transcripts |
| NMD | Nonsense-mediated mRNA decay |
| NGS | Next generation sequencing |
| MAF4 | MADS AFFECTING FLOWERING4 |
| *FLINC* | *FLOWERING LONG INTERGENIC NON CODING RNA* |
| *FLORE* | *CDF5 LONG NONCODING RNA* |
| *PMS1T* | *PHOTOPERIOD-SENSITIVE GENIC MALE STERILITY 1* |
| AP2 | APETALA2 |
| TCV | *Turnip crinkle virus* |
| *ELENA1* | *ELF18-INDUCED LONG-NONCODING RNA1* |
| *TL* | *TWISTED LEAF* |
| PID | PINOID |
| *HID1* | *HIDDEN TREASURE 1* |
| NSR | Nuclear speckle RNA-binding protein |
| *ncW6* | *ncRNA-W6* |
| RACE | Rapid amplification of cDNA ends |
| CAGE | Cap analysis of gene expression |
| PAS-seq | PolyA site sequencing |
| qRT-PCR | quantitative RT-PCR |
| sgRNA | single guide RNA |
| TriFC | Trimolecular fluorescence complementation |
| DHSs | DNase I hypersensitive sites |
| TSI | Tissue specificity index |

| | |
|---|---|
| PCGs | Protein-coding genes |
| DAI | Days after induction |
| STM | SHOOTMERISTEMLESS |
| REV | REVOLUTA |
| VIM3 | VARIANT IN METHYLATION 3 |
| ANT | AINTEGUMENTA |
| AGL24 | AGAMOUS-LIKE 24 |
| PEs | Putative enhancer-like elements |
| la-e | lincRNA-associated enhancers |
| na-e | non-lincRNAs associated enhancers |
| la-g | the neighboring target genes of enhancer associated lincRNAs |
| na-g | the neighboring target genes of non-enhancer associated lincRNAs |
| BRC1 | BRANCHED 1 |
| SUP | SUPERMAN |
| PTL | PETAL LOSS |
| WGD | whole genome duplication |
| CNS | conserved non-coding sequences |
| EAG | Evolutionary age group |
| Ath | *Arabidopsis thaliana* |
| Aly | *Arabidopsis lyrata* |
| Cru | *Capsella rubella* |
| LCM | Laser capture microdissection |
| FACS | Fluorescent activated cell sorting |
| INTACT | Nuclear tagging in specific cell-types |
| RACE-seq | RACE coupled with long-read high-throughput sequencing |
| SCL | SCARECROW-Like |
| CLF | CURLY LEAF |
| PRC | Polycomb Repressive Complex |
| sORFs | small open reading frames |
| RBPs | RNA binding proteins |

# 1. Introduction

Long non-coding RNAs (lncRNAs) are transcripts larger than 200 nucleotides and without protein coding potential. Computational approaches have identified numerous lncRNAs in different plant species. Experimental research has unveiled that lncRNAs participate in a wide range of biological processes, including regulation of flowering time and morphogenesis of reproductive organs, as well as abiotic and biotic stress responses. LncRNAs execute their functionality by interacting with DNA, RNA and protein molecules, and by modulating the expression level of their target genes through epigenetic, transcriptional, post-transcriptional or translational regulation. In the following sections, the characteristics, known functions and molecular mechanisms of plant lncRNAs are summarized. Parts of the introduction were accepted and published as a review in *Planta* and reproduced with permission (Li Chen, Qian-Hao Zhu, and Kerstin Kaufmann. Long non-coding RNAs in plants: emerging modulators of gene activity in development and stress responses. Planta, 2020). The manuscript text and figures were composed and prepared by myself, Dr. Zhu and Prof. Dr. Kaufmann corrected the manuscript and made comments for improvement.

## 1.1 Long non-coding RNAs in plants: emerging modulators of gene activity in development and stress responses

Pervasive transcription of genomes contributes to the large number of non-coding RNAs. Long non-coding RNAs (lncRNAs) are typically defined as transcripts of more than 200 nucleotides length and without any protein coding potential (Quinn and Chang 2016; Budak et al. 2020). Since the discovery of thousands of lncRNAs based on the genome-wide survey, the functional relevance of lncRNAs has been debated. They have been suggested to be 'transcriptional noise' (Hüttenhofer et al. 2005) rather than having specific biological functions (for review, see Kung et al. 2013). It is now becoming clear that lncRNAs represent a highly heterogeneous class of molecules that can be distinguished based on their biogenesis and functions and by their position relative to other genomic features such as protein-coding genes or transposons (Yu et al. 2019a) (**Table 1.1**).

Most lncRNAs are located within intergenic regions although intronic lncRNAs and natural antisense lncRNAs have been reported. Specialized groups of plant lncRNAs produced by RNA polymerase IV or V are important scaffolding components in the RNA directed DNA methylation (RdDM) pathway (Chekanova 2015). Several features of lncRNAs, including transcript length, expression level, and specificity, biogenesis, post-transcriptional processing, and degradation, are not only different from those of protein-coding mRNAs but also heterogeneous among the lncRNAs. Even though large numbers of lncRNAs have been

identified via next generation sequencing (NGS), microarray and comparative genomics, only a small portion of lncRNAs have been functionally characterized. LncRNAs can regulate mRNA expression via *cis* and/or *trans* mechanisms, act as signals and decoys of miRNAs or RNA binding proteins, provide specificity for target molecules such as histone modifying enzymes, and function as scaffolds stitching together large molecular machinery (Wang and Chang 2011). In terms of the layers of regulation, lncRNAs can affect target gene activity at almost all levels of regulation, including chromatin, transcriptional, post-transcriptional, translational, and post-translational levels (Fatica et al. 2014; Lucero et al. 2020). In plants, lncRNAs have been shown to participate in the regulation of developmental processes, biotic and abiotic stress responses, in addition to acting as modulators of the basic cellular machinery. Comparative analysis of lncRNAs in many plant species has deepened our understanding of the conservation and evolution of lncRNAs. Transposable elements contributed significantly to the origin and diversification of lncRNAs in plants (Kapusta et al. 2014). Many identified and experimentally verified lncRNAs have been curated and deposited into databases, making them accessible for functional studies (see, e.g. EVLncRNAs (Zhou et al. 2018, 2019), **Supplemental table S0**). In this review, we summarize the characteristics and recent findings on plant lncRNA functions and document the strategies and experimental approaches used in the identification and analysis of plant lncRNAs.

**Table 1.1: Comparison of typical characteristics of mRNAs and lncRNAs**

| Category | mRNAs | lncRNAs |
|---|---|---|
| Length | Longer | shorter |
| Expression specificity | more constitutive expression | most specifically expressed |
| Expression level | higher expression | lower expression |
| Biogenesis | RNA pol II | RNA pol II, pol III, pol IV, pol V (plant-specific RdDM pathway) |
| TF binding sites | mostly in promoters, regulatory introns, enhancers | promoters and lncRNA gene body |
| Processing | 5’ caps and 3’ polyA tails | most have, some without polyA tails |

### 1.2 Discovery and classification of lncRNAs

The first eukaryotic lncRNA, *H19* with a length of 2.3 kb, was discovered in mouse in 1984 and is highly expressed during embryo development (Pachnis et al. 1984). Both *H19* and its neighboring protein coding gene *Igf2* are imprinted. *H19* and *Igf2* are maternally and paternally expressed, respectively, and form the H19/IGF2 cluster (**Figure 1.1A**) (Nordin et al. 2014; Keniry et al. 2012). Subsequently, many lncRNAs such as *Xist*, *Airn*, *MALAT1,* and *HOTAIR*

were discovered and characterized in animals through genetic, molecular, and functional studies (Fatica et al. 2014). The first identified plant lncRNA, *Enod40*, was isolated as an early marker for nodule organogenesis in *Medicago* plants (Crespi et al. 1994). *Enod40* was found to trigger changes in subcellular localization of the nuclear RNA binding protein MtRBP1 (Crespi et al. 1994; Campalans et al. 2004). Since then, plant lncRNAs have been identified as regulators of miRNA activity (Franco-Zorrilla et al. 2007), epigenetic regulation (Swiezewski et al. 2009; Wu et al. 2020), and modulation of chromatin structure (Ariel et al. 2014, 2020; Kim and Sung 2018). Furthermore, the two antisense lncRNAs *LAIR* (*LRK Antisense Intergenic RNA*) and *MAS (MAF4 antisense RNA)* were found to interact with WDR5 (a component of the COMPASS-like complex) thereby regulating flowering time in rice and *Arabidopsis*, respectively (Wang et al 2019; Zhao et al. 2018).

Based on their genomic position and orientation relative to their neighboring or overlapping protein coding genes, lncRNAs can be classified into intronic lncRNAs, intergenic lncRNAs (lincRNAs), natural antisense lncRNAs, and sense lncRNAs (Ariel et al. 2015, **Figure 1.1B**). LincRNAs can be further classified based on the genomic features with which they are associated, such as promoters, enhancers, and transposable elements (**Figure 1.1).**

Enhancer-associated lncRNAs (eRNAs) are usually less than 2000 nt in length and bidirectionally transcribed from corresponding enhancers, as shown in animal model systems (Shlyueva et al. 2014). These eRNAs often lack polyA tails and are degraded by the exosome when they are released from RNA polymerase II (RNA pol II, Shlyueva et al. 2014). Bidirectional transcripts are not typically detected in enhancers or promoters of *Arabidopsis* and other plants, most likely due to rapid degradation (Thieffry et al. 2020). Most eRNAs are functionally uncharacterized. Data from non-plant model systems suggest roles of eRNAs in mediating changes in chromatin status, though it has also been suggested that they represent products of 'accidental' RNA pol II activity at enhancers (Shlyueva et al. 2014). Transposable element-associated lncRNAs (TE-lncRNAs) overlap with transposons that provide lncRNAs with distinct characteristics and chromatin environment. Transposons such as ALU elements promote nuclear localization of human lncRNAs (Lubelsky and Ulitsky 2018; Carlevaro-Fita et al. 2019). The evolutionary origins and functional diversification of lncRNAs are also influenced by transposable elements (Kapusta et al. 2013). Last but not least, many lncRNAs act as precursors of miRNAs or siRNAs, such as *Iw1* involved in the wax biogenesis of wheat (Huang et al. 2017). Altogether, lncRNAs comprise a highly heterogeneous class of biomolecules that reflect differences in their biogenesis, functionality, and turnover. In the following, we aim to provide an overview on the characteristics of plant lncRNAs, pointing toward their distinct

origins and mechanisms of action.

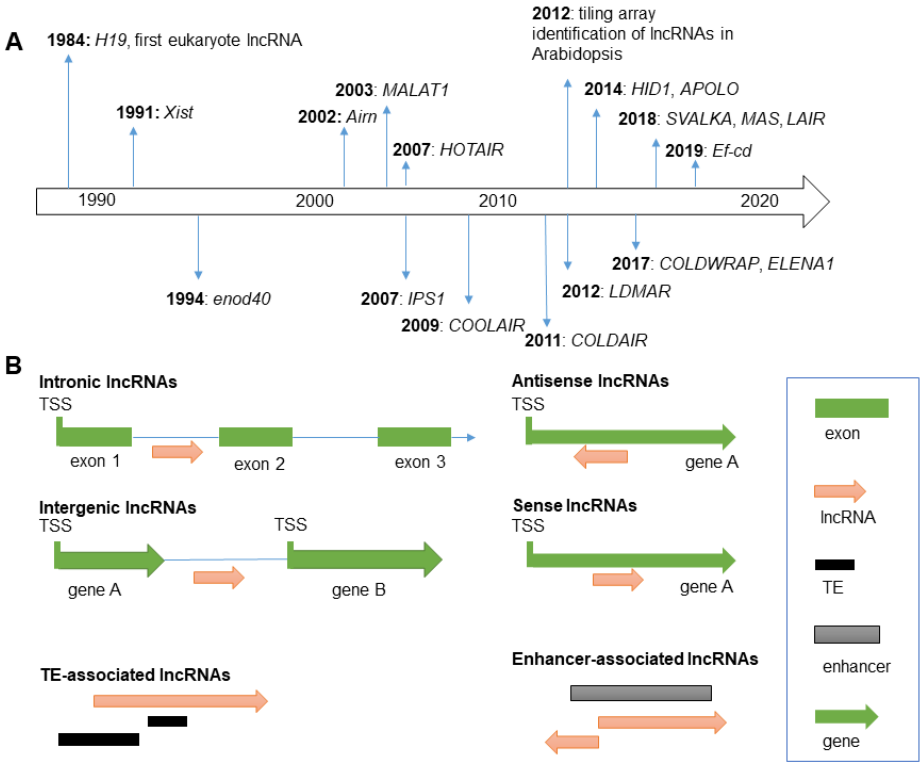



**Figure 1.1: Discovery and classification of lncRNAs**. (A) A timeline of lncRNA discovery. (B) Classification of lincRNAs based on genomic position (enhancer, promoter, the genomic locus of protein-coding genes, transposon (TE)).

## 1.3 Characteristics of lncRNAs

### 1.3.1 Abundance and size of lncRNA transcripts

LncRNAs have been identified in a wide range of plant species including *Arabidopsis*, rice, and maize. The number of lncRNAs identified varies depending on the technology used for identification in each species, and large-scale analyses have reported between 6480 (Liu et al. 2012) and 6510 (Zhao et al. 2018b) lncRNAs in *Arabidopsis* (**Table 1.2**). LncRNAs are usually shorter than protein-coding mRNAs, and they contain fewer exons. Some lncRNAs contain open reading frames (ORFs) with the potential of producing small peptides (Lin et al. 2020). While it is not known whether functional peptides are formed, small ORFs encoded in lncRNAs have been shown to affect the growth of human cells (Chen et al. 2020).

**Table 1.2: Example studies for systematic lncRNA identification in plants.**

| Species | Tissues | Number of lncRNAs | Reference |
|---|---|---|---|
| *Arabidopsis thaliana* | Seedling, inflorescence, | 6,480 | Liu et al. 2012 |
| *Oryza sativa* | anther, pistil, seed, shoot | 2,224 | Zhang et al. 2014 |
| *Brassica rapa* | pollen | 12,051 | Huang 2018 |
| *Gossypium hirsutum* | root, hypocotyl, leaf, flowers | 35,268 | Wang et al. 2015b |
| *Zea mays* | root, leaf, and shoot | 20,163 | Li et al. 2014 |
| *Solanum lycopersicum* | fruits | 3,679 | Zhu et al. 2015a |

### 1.3.2 Expression specificity and functionality

LncRNAs are typically expressed in a more tissue-specific manner than mRNAs of protein-coding genes. In *Arabidopsis*, ~32% of lncRNAs display organ-specific expression that could be verified by experimental methods such as qRT-PCR (Liu et al. 2012). The high expression specificity of lncRNAs makes them potentially suitable as markers for tissues and developmental stages. Partly, the apparent specificity could also be attributed to the generally low expression level of lncRNAs, as well as limitations in detection by standard mRNA-sequencing protocols.

### 1.3.3 Biogenesis, splicing and regulation of lncRNAs

As protein-coding mRNAs, the biogenesis of most lncRNAs depends on RNA pol II-mediated transcription and co-transcriptional splicing. For instance, cold responsive lncRNA *SVALKA* is transcribed by RNA pol II and it tightly regulates expression of *C-REPEAT/DRE BINDING FACTOR 1* (*CBF1*) (Kindgren et al. 2018). Additional factors or other RNA polymerases also contribute to the biogenesis of lncRNAs (Liu et al. 2015). *Arabidopsis* lncRNA *AtR8* is transcribed by RNA pol III and involved in the hypoxic stress response (Wu et al. 2012). A subset of lncRNAs are produced by the plant-specific RNA pol IV or pol V (Liu et al. 2015). These lncRNAs can play a role in the RdDM pathway, in which RNA pol IV-transcribed lncRNAs interact with *INVOLVED IN DE NOVO 2* (*IDN2*) (Zhu et al. 2013). Additionally, components of the miRNA pathway contribute to lncRNA biogenesis. For example, processing of a subset of lincRNAs requires *SERRATE* (*SE*), *CAP BINDING PROTEIN20* (*CBP20*), and *CAP BINDING PROTEIN80* (*CBP80*) (Liu et al. 2012). DICER-like proteins may also play roles in the processing of plant lincRNAs (Ma et al. 2014). Consequently, these plant lncRNAs are usually processed into 24 nt het-siRNA by DCLs (e.g. DCL3) to methylate target genomic loci (e.g. TEs).

During RNA processing, lncRNAs are typically stabilized by capping and polyadenylation in the nucleus. A subset of lncRNAs in mammalians, such as *MALAT1,* are processed by RNase

P, do not possess polyA tails and, instead, have a specialized 3' end structure (Wilusz et al. 2008). In humans, non-polyadenylated lncRNAs (i.e. sno-lncRNAs) that are flanked by snoRNAs and protected by RNA binding proteins have also been identified (Yin et al. 2012). Among the non-polyadenylated lncRNAs, a specialized form of RNAs called circRNAs, such as *circSEP3* in *Arabidopsis* (Conn et al. 2017), join their heads with tails covalently in a process called back-splicing that is mediated by the spliceosome machinery (Chen, 2016). CircRNAs may regulate the splicing of their cognate mRNAs, as was shown for circSEP3 and its target *SEPALLATA3* (*SEP3*) (Conn et al. 2017). Differential polyadenylation, linked with changes in preferential subcellular localization, in response to stress, has been described for rice and *Arabidopsis* lncRNAs(Di et al. 2014; Yuan et al. 2016, 2018).

In mammalians, ~13% of lncRNAs are transcripts that are derived from divergent transcription in promoters of protein-coding genes (Grzechnik et al. 2014). These divergent transcripts are associated with histone modification (e.g. H3K56ac), RNA pol II Tyr1 phosphorylation, and chromatin remodeling factors (e.g. SWI/SNF). Furthermore, the directionality of these divergent lncRNAs is determined by the asymmetry of U1 snRNP and polyadenylation signals (Quinn and Chang 2016). However, divergent transcription does not appear to occur in the majority of genes in *Arabidopsis thaliana* (Thieffry et al. 2020a; Hetzel et al. 2016). In addition to the RNA polymerase machinery, transcription factors (TFs) and chromatin environment (e.g. histone modification and DNA methylation) also contribute to the regulation of lncRNA expression (Quinn and Chang 2016).

Data from humans suggest that the splicing efficiency of lncRNAs is lower than that of mRNAs, possibly due to lower binding of splicing factors and the presence of weaker splicing-related motifs (Melé et al. 2017). Low sequencing depth and limitation of RNA-seq assembly methods may also contribute to this observation since RACE-seq of lncRNAs detected as many alternative splicing events in lncRNAs as in mRNAs (Lagarde et al. 2016).

### 1.3.4 Structure of lncRNAs

LncRNAs possess secondary structures which may be necessary for their functionality. There are usually two types of functional sites in lncRNAs: interacting sites which are necessary for sequence-specific interactions with RNA binding proteins, and structural sites which confer the identity of secondary and/or tertiary structures directing interacting partners (Fabbri et al. 2019). For example, *COOLAIR* participating in vernalization has a multi-way junction motif and two right-hand turn motifs (Hawkes et al. 2016), which are very conserved secondary structures in the *Brassicaceae* family. However, it is still unknown which proteins interact specifically with these motifs.

### 1.3.5 Subcellular localization of lncRNAs

mRNAs are usually exported into the cytosol for translation. By contrast, after processing lncRNAs can reside in the nucleus or get exported to the cytosol or other subcellular locations and organelles, such as mitochondria, as demonstrated by RNA FISH and ribosome profiling (Carlevaro-fita and Johnson 2019). Data from animal model systems showed that lncRNAs are generally prone to be more enriched in the nucleus than in the cytoplasm compared to mRNAs (Derrien et al. 2012). Sequence elements within lncRNAs as well as RNA binding proteins contribute to the nuclear or cytosolic localization of lncRNAs, which reflects their cellular roles and functionality (Carlevaro-fita and Johnson 2019). For example, human lncRNAs containing ALU repeats are more prone to be retained in the nucleus because of the binding of specific splicing factors such as HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN K (HNRNPK; Lubelsky and Ulitsky 2018). Some cytosolic lncRNAs are associated with mono- and poly-ribosomal complexes (Bazin et al. 2017; Hsu et al. 2016), and some of these lncRNAs could eventually contribute to the biogenesis of small peptides. A set of nuclear lncRNAs are bound by chromatin, and this localization can be stabilized by U1 snRNP (U1 small nuclear ribonucleoprotein particle) in mammals (Yin et al. 2020). Chromatin-associated lncRNAs potentially influence TF binding or the functionality of enhancers (Shlyueva et al. 2014). While these data from animal model systems indicate intricate mechanisms underlying the subcellular distribution of lncRNAs, less is known on plant lncRNAs. Many identified lncRNAs (e.g. *COOLAIR*, *DRIR*) in plants are localized to and act in the nucleus. For example, cold induced *COOLAIR* coats the *FLC* locus in the nucleus and acts in *FLC* repression by changing the histone modification status (e.g. H3K36me3) dynamics (Rosa et al. 2016; Wu et al. 2020). On the other hand, there are also cytoplasm localized cis-Natural Antisense Transcripts (cis-NATs) overlapping with protein coding genes and some of them could impact the translation of mRNAs (Deforges et al. 2019). In sum, the different types of subcellular localization suggest various molecular mechanisms of action of lncRNAs in transcriptional and posttranscriptional control of gene expression as well as genomic regulation.

### 1.3.6 Decay of lncRNAs

In terms of turn-over of lncRNAs, the half-lives of lncRNAs are typically shorter than those of mRNAs, which reveals complex regulation of lncRNA metabolism in plants (Szabo et al. 2020). LncRNAs are less efficiently synthesized and rapidly degraded (Mukherjee et al. 2017). Like mRNAs, plant lncRNAs can be degraded by both 3’-5’ exonucleolysis via the nuclear exosome and 5’-3’ exonucleolysis via exonucleases such as XRN2 and XRN3 (Kurihara et al. 2012). In mutants of exosome subunits, a set of specialized lncRNAs similar to CUTs (Cryptic

unstable transcripts) and PROMPTs (Promoter upstream transcripts) emerged from TSSs of mRNAs (Chekanova 2015; Chekanova et al. 2007; Thieffry et al. 2020a). Data from humans suggest that exosome-regulated lncRNAs modulate the activity of enhancers, resolving deleterious R-loop structures by the exosome (Pefanis et al. 2015; Nair et al. 2020). Similar to mRNAs, the quality of plant lncRNAs is also surveilled by the non-sense-mediated mRNA decay (NMD) pathway (Drechsel et al. 2013; Kurihara et al. 2009; Kirn et al. 2009). Interestingly, the *up-frameshift* (*upf)* mutants, defective in a component of the NMD pathway, accumulate high levels of transcripts derived from antisense transcription and intergenic regions (Kurihara et al. 2009). This suggests extensive regulation of lncRNA stability via several molecular regulatory pathways.

**1.4 Functionality and molecular mechanisms of lncRNAs in plants**

The recently established lncRNA database EVLncRNAs collected 1543 experimentally validated lncRNAs from 77 species, including 428 lncRNAs from 44 plant species such as *Arabidopsis* and rice (Zhou et al. 2018, 2019). Despite the limited functional characterization of most lncRNAs, studies so far have uncovered a wide range of possible functions and molecular mechanisms mediated by plant lncRNA activities (Datta and Paul 2019) (**Figure 1.2A**).

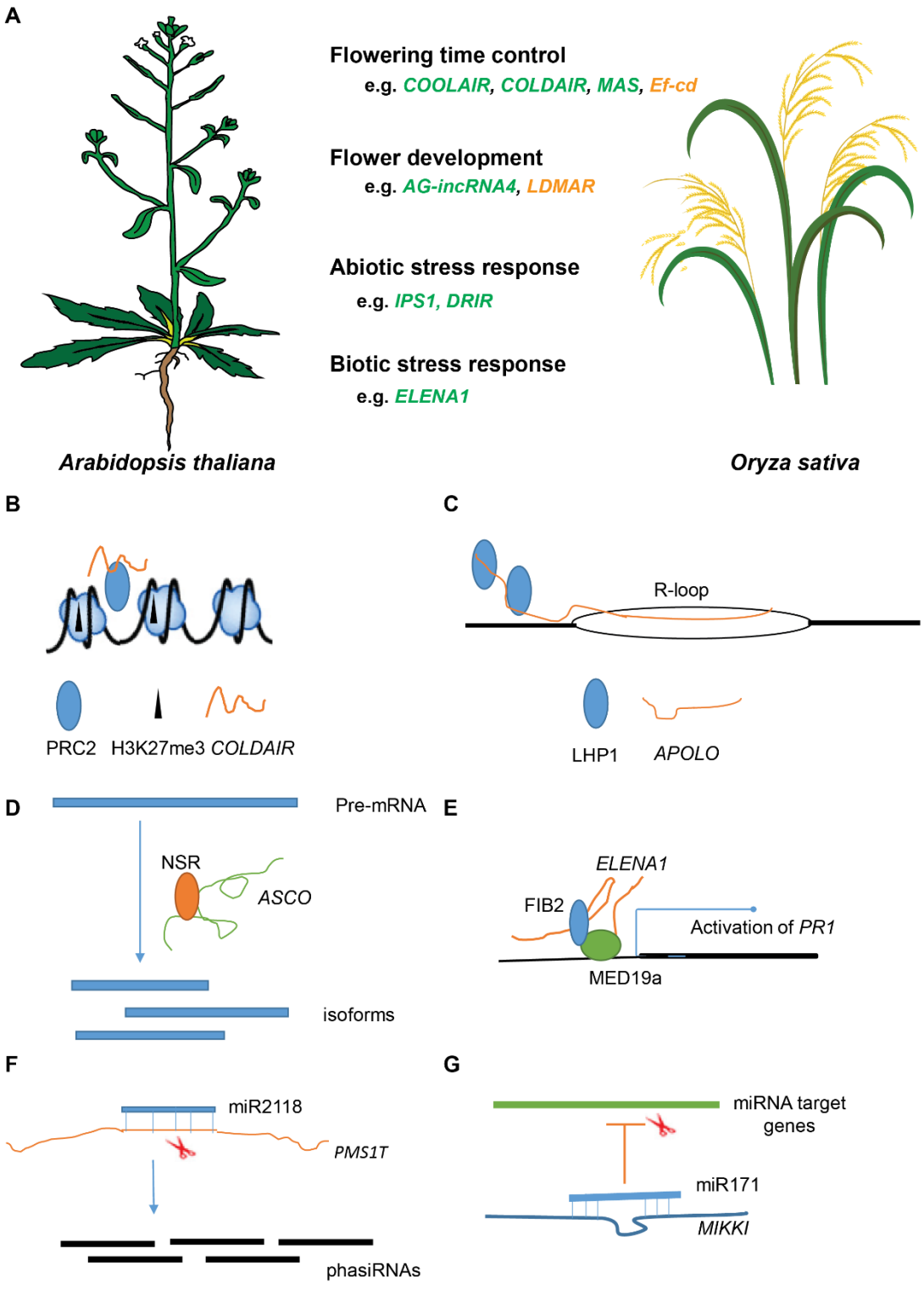



**Figure 1.2**: **Functions of lncRNAs in plants.** (A) LncRNAs participate in diverse biological processes, including flowering time control, flower development, abiotic and biotic stress responses (lncRNAs of *Arabidopsis thaliana* and *Oryza sativa* are highlighted in green and orange, respectively). Illustrations of *Arabidopsis thaliana* and *Oryza sativa* plant are from (Illustrations 2017). (B) *COLDAIR* recruits PRC2 complex to deposit H3K27me3 marks at target gene *FLC* and thereby drives repression of *FLC*. (C) *APOLO* recognizes target gene by R-loop formation and decoys PRC1 protein. (D) *ASCO* can hijack

splicing factor NSR to regulate alternative splicing of target genes. (E) *ELENA1* evicts FIB2 from the FIB2-MED19a complex and contributes to the activation of *PATHOGENESIS-RELATED GENE 1* (*PR1*). (F) miR2118 targets *PM1T* to produce many phasiRNAs. (G) *MIKKI* acts as a target mimic to sequester miR171 away from its target.

### 1.4.1 Regulation of flowering time

Reproductive success in plants is tightly coupled to the proper timing of the floral transition and to robust flower morphogenesis. Flowering time control in plants is regulated via internal signals such as plant hormones and environmental cues including day length and temperature. For *Arabidopsis*, a prolonged period of cold (winter) downregulates in a process called vernalization the expression of the major flowering repressor *FLOWERING LOCUS C* (*FLC)* to promote flowering in spring. There are several lncRNAs intricately and tightly fine-tuning the expression level of *FLC*, such as *COOLAIR*, *COLDAIR*, *ANTISENSE LONG* (*ASL)* and *COLDWRAP* (Hawkes et al. 2016; Castaings et al. 2014; Swiezewski et al. 2009; Rosa et al. 2016; Kim and Sung 2018; Heo and Sung 2011; Csorba et al. 2014; Kim et al. 2017; Shin and Chekanova 2014). *COOLAIR,* including two short and long isoforms with polyA tails, is a class of natural antisense transcripts originating from the 3' end of the *FLC* locus (Swiezewski et al. 2009). *COOLAIR* activity is regulated by 3' processing factors *FCA, FY, FPA, CstF64,* and *CstF77* (polyadenylation cleavage factors), and *PRP8* (the spliceosome component) (Liu et al. 2010; Marquardt et al. 2014). However, detailed molecular mechanisms of *COOLAIR* repressing *FLC* are still unknown, although the increasing level of histone demethylase FLD has been shown to contribute to H3K4me2 demethylation of *FLC* (for review, see Wu et al. 2020). *COLDAIR* is transcribed from the second *FLC* intron and acts as a signal of early vernalization by recruiting the H3K27me3 writer CURLY LEAF (CLF), an enzymatic component of the PRC2 complex and a homolog of EZH2 in animals to repress *FLC* (**Figure 1.2B**) (Kim et al. 2017; Heo and Sung 2011). *COLDWRAP* is a lncRNA associated with the promoter of *FLC*, which also interacts with CLF to form an intragenic chromatin loop and to confer *FLC* repression (Kim and Sung 2018). Furthermore, a non-polyadenylated antisense transcript (ASL, for Antisense Long) is produced from the FLC locus. The function of *ASL* is still unknown but the expression level of *ASL* is downregulated in an *rrp6l* mutant (one of the exosome components, *rrp6l1 rrp6l2* double mutant) (Shin and Chekanova 2014). *MAS* (*NAT-lncRNA_2962*) is a natural antisense lncRNA from the *MADS AFFECTING FLOWERING4* (*MAF4)* locus involved in vernalization and regulates *MAF4* via interacting with histone modifying enzyme WDR5a (Zhao et al. 2018b).

Other flowering time-related lncRNAs, including *FLOWERING LONG INTERGENIC NON CODING RNA* (*FLINC*), *CDF5 LONG NONCODING RNA* (*FLORE*), *LDMAR*, *PHOTOPERIOD-*

*SENSITIVE GENIC MALE STERILITY 1* (*PMS1T*), and *Ef-cd*, have been recently discovered in *Arabidopsis* or rice (Severing et al. 2018; Henriques et al. 2017; Ding et al. 2012a, 2012b; Fan et al. 2016; Fang et al. 2019). *FLINC* regulates ambient temperature mediated flowering. T-DNA insertion mutants of *FLINC* flowered earlier due to upregulated *FT* expression while the underlying mechanism is not known (Severing et al. 2018). The circadian-regulated *FLORE* is a lncRNA antisense to *CDF5* and is involved in promoting photoperiodic flowering by repression of several *CDFs* and consequently activation of *FT* (Henriques et al. 2017). In sum, the different examples indicate interesting functions for lncRNAs in the environment-dependent modulation of flowering time, providing model systems for studying how gradual changes in environmental factors trigger a defined developmental decision at the transcriptional or posttranscriptional level.

### 1.4.2 Modulation of reproductive organ development

After the floral transition, the inflorescence meristem starts to produce floral meristems, which in turn give rise to different types of floral organs. Nowadays, a number of lncRNAs such as *LINC-AP2* (Gao et al. 2016), *LONG-DAY SPECIFIC MALE-FERTILITY-ASSOCIATED RNA (LDMAR)* (Ding et al. 2012a,b), *PHOTOPERIOD-SENSITIVE GENIC MALE STERILITY T* (*PMS1T,* Fan et al. 2016), and *EARLY FLOWERING-COMPLETELY DOMINANT* (*Ef-cd;* Fang et al. 2019), have been found to regulate diverse aspects of flower and reproductive development (see **Supplemental table S0** for a more comprehensive list of examples). *LINC-AP2* is an intergenic lincRNA close to the flower developmental regulatory TF gene *APETALA2* (*AP2)*. While *AP2* is downregulated upon infection with *Turnip crinkle virus* (TCV), the expression of *LINC-AP2* is elevated, and strong upregulation of *LINC-AP2* correlates with abnormal floral structures (Gao et al. 2016). The long intergenic rice lncRNA *XLOC_057324* is highly expressed in reproductive organs, and T-DNA insertion mutant analysis suggests roles in the control of flowering and plant fertility(Zhang et al. 2014).

Other functions of lincRNAs include processes directly related to plant fertility. *BcMF11* is specifically expressed in pollen and is necessary for male fertility and pollen development in *Brassica campestris ssp*. chinensis (Song et al. 2013). *SUPPRESSOR OF FEMINIZATION* (*SUF)* is a lncRNA antisense to *MpFGMYB*, an important regulator of female sexual tissue differentiation in liverwort (*Marchantia polymorpha*). The *suf* loss of function mutant created by Cas9-based deletion displayed male-to-female sexual conversion, probably due to failure to repress *MpFGMYB* in male tissues in the absence of *SUF* (Hisanaga et al. 2019). The intronic lncRNA *AG-incRNA4* from the second intron of the floral homeotic *AGAMOUS* (*AG*) gene in *Arabidopsis* is expressed in leaves and interacts with the PRC2 complex component CLF to

deposit H3K27me3 histone marks onto the *AG* locus, thereby contributing to repression of *AG* expression in leaves. The knockdown of *AG-incRNA4* resulted in *AG* activation in leaves by lowering the H3K27me3 level at the *AG* locus. Consequently, the corresponding mutant showed phenotypes resembling those of ectopic *AG* expression (Wu et al. 2018). *LDMAR* was identified in rice through map-based cloning and regulates photoperiod-sensitive male fertility via RdDM (Ding et al. 2012a, 2012b; Zhou et al. 2012).

Small RNAs, including het-siRNAs, phase-siRNAs and miRNAs play a critical role in development and stress responses. For example, miR396-mediated regulation of *HaWRKY6* plays a role in the protection of damage caused by high temperature in sunflower and affects plant growth (Giacomelli et al. 2012). The identification of ncRNA-W6 (*ncW6*) in the promoter of *HaWRKY6* revealed another layer of regulation of the gene by a nonNSR-coding RNA. *ncW6* derives from a transposon of the MITE family and can form a hairpin structure that is processed into 24 nt het-siRNAs by DCL3 to trigger DNA methylation in the flanking regions of *HaWRKY6*. DNA methylation changes the chromatin structure of the *HaWRKY6* locus and promotes the formation of a loop encompassing the whole locus to enhance transcription of *HaWRKY6*. The level of DNA methylation, and consequently the formation of the loop and the expression level of *HaWRKY6,* is regulated in a tissue-specific manner (Gagliardi et al. 2019). Another lncRNA, *PMS1T*, identified by map-based cloning in rice, contributes to photoperiod-sensitive male sterility by producing phase-siRNAs in a miR2118-dependent manner (Fan et al. 2016) (**Figure 1.2F**). *Ef-cd* is an antisense RNA in the *OsSOC1* locus and positively regulates *OsSOC1* activity by deposition of H3K36me3, thereby reducing the time-span that is needed to reach plant maturity without yield penalty (Fang et al. 2019). Together, these findings highlight important functions for lncRNAs in reproductive growth via different molecular mechanisms. Since many uncharacterized lncRNAs are associated with genomic loci that encode developmental control genes, these will provide interesting targets for future research.

### 1.4.3 Response to abiotic and biotic stresses

As sessile organisms, plants must cope with various kinds of abiotic and biotic challenges. Plants have evolved intricate signaling cascades and molecular networks to combat these stresses. Under phosphate starvation conditions, *Arabidopsis* plants express the lncRNA *Induced by Phosphate Starvation 1* (*IPS1*). *IPS1* acts as an endogenous target mimic to sequester and repress miR399, a repressor of *PHOSPHATE2* (*PHO2)*, which encodes a ubiquitin-conjugating E2 enzyme. Repression of *PHO2* enhances phosphate uptake and accumulation (**Figure 1.2G**) (Franco-Zorrilla et al. 2007). *ELF18-INDUCED LONG-NONCODING*

*RNA1* (*ELENA1*) is a 589-nt lincRNA conferring immunity of *Arabidopsis*. Plants with a reduced expression level of *ELENA1* by an artificial miRNA are more sensitive to the bacterial pathogen *Pseudomonas syringae pv. tomato* DC3000 and show downregulation of several immunity marker genes, including *PATHOGENESIS-RELATED GENE 1* (*PR1*). In contrast, overexpression of *ELENA1* activates immune genes such as *PR1*. *ELENA1* exerts its role via interacting with components of Mediator (**Figure 1.2E**) (Seo et al. 2017). The lncRNA *DROUGHT INDUCED LNCRNA* (*DRIR)* in *Arabidopsis* positively regulates salt and drought response. Plants overexpressing *DRIR* showed enhanced salt and drought tolerance and displayed higher survival rates under salt and drought stress conditions (Qin et al. 2017). Many other stress response-related lncRNAs have been identified, but their molecular mechanisms of action are yet to be investigated (see, e.g. Zhu et al. 2014; Wang et al. 2017b).

### 1.4.4 Functions in other biological processes

LncRNAs have been shown to participate in diverse biological processes, such as leaf development, auxin signaling, and photomorphogenesis. *TWISTED LEAF* (*TL*) is a rice lncRNA antisense to *OsMYB60* and required for maintaining leaf blade flattening by regulating the expression of its sense mRNA (Liu et al. 2018). The auxin responsive *Arabidopsis* lncRNA *APOLO* plays a role in fine-turning the transcription of its neighboring *PINOID* (*PID*) gene, an important regulator of auxin polar transport, via the formation of a chromatin loop involving the promoter of *PID*. The expression level of *APOLO* determines the chromatin environment in the promoter region of *PID* affecting histone modifications and the level of DNA methylation, and consequently the formation of the chromatin loop and the expression level of *PID* (**Figure 1.2C**) (Ariel et al. 2014). In addition to these *cis* effects, *APOLO* also regulates target loci in *trans* by the formation of R-loop (DNA-RNA duplexes) mediated by short sequence complementarity and thereby decoying PRC1 to target loci to modulate their chromatin status (Ariel et al. 2020). Furthermore, the photomorphogenesis-related lncRNA *HID1* (*HIDDEN TREASURE1*) represses the transcriptional activity of its target gene *PHYTOCHROME INTERACTING FACTOR 3* (*PIF3).* HID1 forms a large nuclear complex with as yet unknown proteins and modulates the chromatin structure in the *PIF3* promoter, consequently repressing hypocotyl elongation of *Arabidopsis* seedlings (Wang et al. 2014b).

LncRNAs function in basic nuclear regulatory processes by interacting with proteins. For example, nuclear speckles are nuclear domains enriched with splicing related factors and located in interchromatin regions of nucleoplasm (Spector and Lamond 2011). It was shown that *Arabidopsis ASCO-lncRNA* competes for the NUCLEAR SPECKLE RNA-binding proteins (NSRs) and sequesters NSRs to modulate the alternative splicing pattern of target genes

(**Figure 1.2D**) (Bardou et al. 2014). LncRNAs are also components of the telomerase molecular machinery. For example, lncRNA *AtTR* is the RNA subunit of telomerase, which interacts with *TELOMERASE REVERSE TRANSCRIPTASE* (*TERT*) to maintain the integrity and stability of telomeres (Song et al. 2019; Michal et al. 2019). This indicates roles of lncRNAs in genome integrity and genome functions beyond biological functions in development or environmental response, which emphasize the need for multiscale experimental methodologies to characterize lncRNA functions.

### 1.5 Experimental methodologies for functional characterization of lncRNAs

Similar to protein-coding genes, functions of lncRNAs can be investigated using forward and reverse genetics approaches. However, functional analysis of lncRNAs is hampered by the need to distinguish the functions of the lncRNA transcript from that of its genomic locus. This is because lncRNAs are often produced from DNA genomic regions with other functions, e.g. loci of protein coding genes (in the case of intronic or antisense lncRNAs) or enhancers (e.g. in the case of eRNAs). In addition, RNAi-based knockdown of lincRNA activities can have side effects that are not related to the functions of lincRNAs, for instance, RNAi-mediated DNA methylation is possible to change the functionality of the genomic regions in other aspects (e.g. affecting enhancer activity). Finally, not the lincRNA transcript itself, but the process of transcription may exert a regulatory function (Gowthaman et al. 2020).

In plants, a small set of lncRNAs has been identified by map-based cloning and functionally characterized, such as *LDMAR* (Ding et al. 2012a), *PMS1T* (Fan et al. 2016), *Ef-cd* (Fang et al. 2019), and *Iw1* (Huang et al. 2017). However, reverse genetics (e.g. based on T-DNA mutagenesis populations, RNAi, overexpression) is most commonly used for studies of lncRNA functions, because the vast majority of lncRNAs were identified by high throughput technologies. Every method used to perturb lncRNA functions has disadvantages. For example, T-DNA insertions or CRISPR/Cas9-based deletions in intergenic regions may not only inhibit lncRNA expression but also affect other functions of the DNA sequences, such as TF binding sites or regulatory elements within lincRNA loci, thereby altering the expression of nearby protein coding genes. When studying antisense, sense or intronic lncRNAs, these approaches can also have side effects, such as modifying splicing of the associated protein-coding genes. The RNAi technology on the other hand is known to be prone to off-targeting and may cause RdDM, thereby confounding functional interpretation of the target lncRNAs. Thus, a combination of different approaches and proper control experiments are required to study lncRNA functions.

Here we propose a workflow for functional investigation of plant lncRNAs (**Figure 1.3**).

When a candidate lncRNA is identified, the first task to perform a comprehensive inspection of the sequence and structure of the lncRNA. Rapid amplification of cDNA ends (RACE) can be used to obtain the full length transcript(s) of the lncRNA. Searching publicly available datasets, such as cap analysis of gene expression (CAGE) and polyA site sequencing (PAS-seq) (Shepard et al. 2011), and performing RNA-seq will give clues about the general structure as well as alternative splicing patterns of the lncRNA locus of interest. Northern blotting and quantitative RT-PCR (qRT-PCR) are standard approaches for the investigation of the expression profiles of lncRNAs. GREEN FLUORESCENT PROTEIN (GFP) reporter imaging can be used to study dynamic lncRNA promoter activity. RNA-FISH allows the study of the activity and localization of lncRNAs to the level of individual genomic loci (Rosa et al. 2016). Recent studies showed that some lncRNAs could translate into small peptides, and thus it is necessary to distinguish whether the lncRNA of interest functions as non-coding RNA or as small peptide. Several bioinformatics and experimental approaches can be employed for this purpose, such as CPC2 to test coding potential test (Kang et al. 2017). Additionally, lncRNAs should be queried in protein databases including Pfam (Finn et al. 2016) and Uniprot (The UniProt Consortium 2017) to know whether they have potential homologous proteins. Ribosome footprints based on ribosome profiling are indicative of open reading frames, which are used to discriminate lncRNAs from protein coding genes (Lander 2014; Hsu et al. 2016; Bazin et al. 2017). Loss/gain-of-function mutants are generated to investigate the functionality of the lncRNA. Since every technique has its own limitations (see above), it is necessary to use multiple different approaches such as T-DNA mutagenesis, RNAi, overexpression with constitutive and tissue-specific promoters, and CRISPR/Cas9-based mutagenesis combined with mutant complementation. A large number of publicly available T-DNA insertion lines are available for both *Arabidopsis* and rice. Analysis of independent mutant alleles and, importantly, transgenic mutant complementation (in *trans*) can be used to validate the functionality of lncRNAs (see, e.g. Fang et al. 2019). When a lncRNA has multiple isoforms, generating mutants for each isoform can distinguish the roles of individual isoforms. CRISPR/Cas9-based mutagenesis usually creates small indels in the target site (Li et al. 2018), which might not influence the functionality of the lncRNA. This can be overcome by introducing a pair of single guide RNA (sgRNA) to induce a larger indel in the corresponding lncRNA locus. The use of multiple such pairs of sgRNAs covering the entire lncRNA can help to dissect the functional regulatory sites of the lncRNA. In these experiments, potential side effects arise from mutagenizing other functional DNA elements that reside within the lncRNA locus. Therefore, the target lncRNA locus should be evaluated carefully by taking into account

existing information on TF binding sites or chromatin structure. In all types of mutant analyses, the phenotypic analyses should be complemented by monitoring changes in expression of the protein-coding genes flanking the lncRNA locus of interest. Especially for studying *trans* mechanisms of lncRNAs, (inducible) ectopic expression or artificial miRNA technology can be used for validation.

Functional lncRNAs typically interact with DNA, RNA and proteins. The *in vitro* or *in vivo* approaches developed for investigating the RNA-protein (e.g. RIP and CLIP)(Cao et al. 2019), RNA-DNA (e.g. ChIRP) (Chu et al. 2012), and RNA-RNA (e.g. RAP-RNA) (Engreitz et al. 2014) interactions can be used to identify the molecular partner(s) interacting with lncRNAs. The subcellular localization of lncRNAs is also important, since it may provide clues on functions. For example, single molecule RNA FISH analysis revealed that *COOLAIR* and *FLC* transcripts are mutually exclusively expressed (Rosa et al. 2016). It is important to further develop *in vitro* and *in vivo* experimental methods to screen and validate the interaction between lncRNAs and their partner molecules. For example, a trimolecular fluorescence complementation (TriFC) system has been used to demonstrate lncRNA-protein interaction by tagging a lncRNA with the MS2 system (MS2 sequence and phage MS2 coat protein fused to YFP-N) and co-transfecting it together with YFP-C tagged RNA-binding protein into tobacco leaves via *Agrobacterium* (Seo et al. 2019). Finally, we envision that efficient novel experimental and computational methods will be developed for investigation of the functionality of lncRNAs in plants at the level of single cells or subcellular compartments.

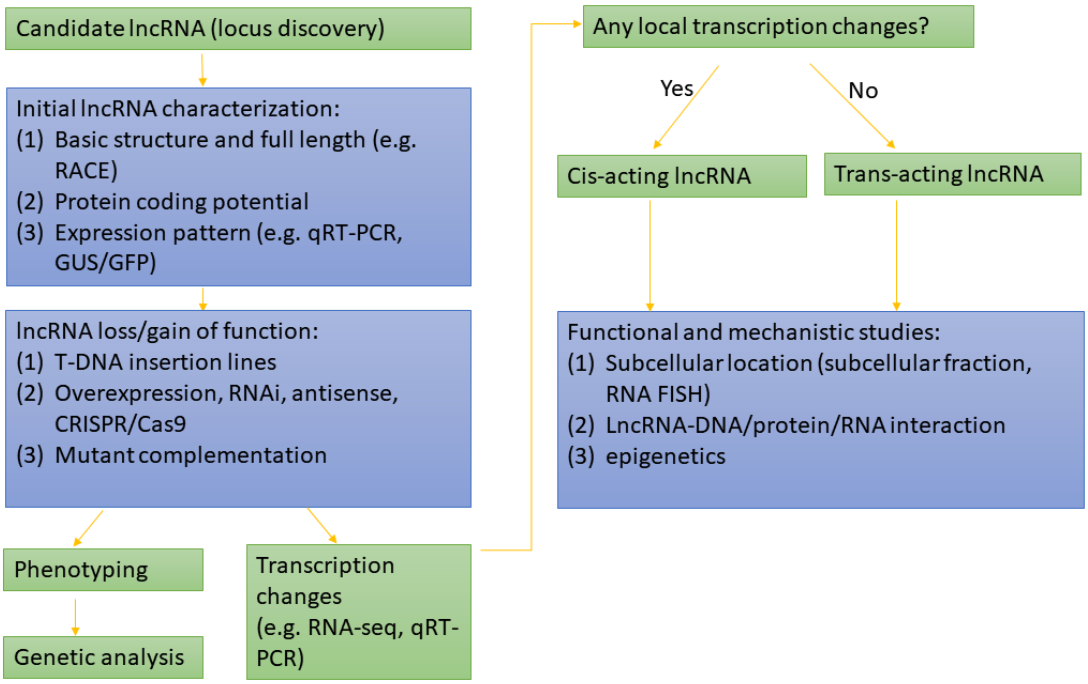


**Figure 1.3**: Experimental workflow for dissection of lncRNA functions. Details are described in the main text.

### 1.6 Aims of the project

The project aims to deepen the understanding of lncRNAs with plants as the model system. Previously, plant lncRNAs were demonstrated to function epigenetically to modulate the expression and functionality of their target genes by modifying histone and chromatin modifications. However, our understanding of the molecular mechanisms of plant lncRNAs is still elusive. Additionally, for a specific biological context, the features and functions of lncRNAs need to be elucidated. For instance, the molecular mechanism of lncRNAs participation in the floral gene regulatory network remains to be unknown. Moreover, evolutionary studies of lncRNAs by comparative genomes could facilitate and prioritize primary sequence associated functions of lncRNAs and pinpoint critical regions of lncRNAs. However, the evolutionary dynamics of lncRNAs in plants including non-flowering plants still remain to be elusive and the expression pattern in different plant species was quite unknown. Therefore, in this study, we intend to answer the above questions.

**In the first part of the study**, flower developmental time-series transcriptome data are used to systematically identify thousands of flower-related long intergenic RNAs (lincRNAs). These lincRNAs are investigated to elucidate positional, chromatin, and structural features, and to predict functions and mechanism in the floral gene regulatory network.

Besides focusing on the model plant *Arabidopsis thaliana*, the analysis of evolutionary conservation can provide important insights into the relevance and potential roles of lncRNAs. The availability of high-quality plant genomes along with rich genomic resources, such as transcriptomes of different developmental stages in various plant species, makes this a highly timely topic. **In the second part of the study**, thousands of lncRNAs are identified using a standardized pipeline (Kapusta et al. 2014) in 26 plant species including non-flowering plants, allowing to infer homologous lncRNAs and explore conserved characteristics and functions of lncRNAs during land plant evolution.

## 2. Materials and methods

### 2.1 LincRNAs identification and expression quantification

We used a developmental time series dataset in *Arabidopsis thaliana* from the AP1-based floral synchronized system generated at 0, 2, 4, 8 days after induction (Chen et al. 2018). Additional RNA-seq datasets in NCBI SRA and GEO database were selected by focusing on flowers and meristem samples in *Arabidopsis* (**Supplemental Table S1**). The raw datasets of paired-end/single-end RNA-seq reads (details described in **Supplemental Table S1**) were used for quality to filter low-quality reads, adaptor sequences using software FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). The lincRNA dataset was assembled using a reference guide transcriptome assembly pipeline (shown in **Figure 3.1B).** Briefly**,** filtered clean reads of RNA-seq were mapped to the *Arabidopsis thaliana* genome (TAIR10, https://www.arabidopsis.org/) with hisat2 (Kim et al. 2015) mapping software with default parameters. The mapped reads were used for assembling transcripts of each sample (replicate or experiment) separately with StringTie (Pertea et al. 2015) and StringTie merge module (stringtie --merge) were utilized to obtain merged transcripts for all RNA-seq samples (details described in **Supplemental Table S1**, mainly including meristem and floral samples). For merged transcripts, we compared it with the reference TAIR10 annotation (https://www.arabidopsis.org/) and Araport annotation (https://www.araport.org/) to filter out known PCGs using the gffcompare module of stringtie, and to filter out transcripts of less 200nt transcripts that have a predicted putative longest ORF peptide of more than 100 amino acids, as well as RNAs that have hits from protein databases (Nr, UniprotKB and Pfam) using Blast search. We also excluded other types of non-coding RNAs such as miRNAs, rRNAs, tRNAs, snoRNAs, snRNAs etc from the transcripts. Portions of transcripts from the assembled and merged transcripts with StringTie are marked with unknown strand information which was filtered by the pipeline in order to retain high confident transcripts. The remaining transcripts were evaluated using the coding potential software COME (Hu et al. 2017). The final set of potential lncRNAs was classified into lincRNAs, intronic lncRNAs, and antisense lncRNAs. The expression value of all genes loci including PCGs and lincRNAs are determined using StringTie (Pertea et al. 2015) and ballgown (Frazee et al. 2015) software. If samples consisted of replicates, the expression value can be obtained using the average value of these replicate samples. We only retained genes whose expression value is more than 0.5 TPM as “expressed gene loci” in at least one of the datasets (PCGs and lincRNAs) to discard transcription artifacts. Owing to a subset of lincRNAs lacking polyA tails, total RNA-seq libraries from the same developmental time points described above were also generated in the AP1-GR based floral induction system to identify lincRNAs (SRA accession number: PRJNA610830). With the same

pipeline used before, 616 lincRNAs were identified, of which ~77% were also identified in the polyA-RNA-seq data. The RNA-seq data were generated by Johanna Müschner (Kaufmann lab).

We selected 25 plant species except for *Arabidopsis thaliana* in the identification of lincRNAs for studying the evolution of lincRNAs. For each plant species, we collected RNA-seq datasets from 12 (Aal) to 899 (Zma) different samples (**Supplemental Table S16**) from the public databases (e.g. NCBI SRA and EBI ENA). Due to most RNA-seq samples without strand information, it is difficult to distinguish antisense RNAs from its parent protein coding genes and thus here we only focus on long intergenic non-coding RNAs (lincRNAs). To identify lincRNAs, the RNA-seq reads were mapped to each plant genome by hisat2 (Kim et al. 2019) with default parameters. The mapped reads of each replicate/sample were assembled by stringtie (Pertea et al. 2015) to get assembled transcripts which were merged together with the stringtie merge module to obtain merged transcripts for all RNA-seq samples of each plant species. The merged transcripts of each plant species were compared with the annotated protein coding genes (PCGs) to filter out protein coding genes by the stringtie gffcompare module. The remaining transcripts were further filtered to remove transcripts less than 200nt, the predicted longest ORF encoding a peptide less than 100 amino acids, or similarity with the protein database NR. The assembled transcripts were queried by blastx and transcripts with E $<10e^{-10}$ was considered to be putative coding transcripts. The remaining transcripts of the last step were evaluated with the coding potential software CPC2 (Kang et al. 2017) to get final long non-coding transcripts which were classified into long intergenic non-coding RNAs (lincRNAs), intronic RNAs, and antisense lncRNAs according to the genomic position of lncRNAs with PCGs. The final lncRNAs and PCGs were merged together to obtain the reference transcripts used for index with Kallisto (Bray et al. 2016). Expression levels of both lincRNAs and PCGs were determined by Kallisto (Bray et al. 2016) with default parameters. If samples had replicates available, the expression levels were estimated based on the average value of the replicated samples. We only retained lincRNAs and PCGs with a TPM > 0.5 to filter out transcriptional noises.

**2.2 Differential expression of lincRNAs and PCGs in *Arabidopsis thaliana***

The raw read counts for all loci of each dataset were estimated using StringTie (Kim et al. 2016) and ballgown (Frazee et al. 2015) software. The raw read count matrix was used as the input for differential expression analysis with the software DESeq2 (Love et al. 2014) and those gene loci including lincRNAs and PCGs with criteria of log2FC>= 1 or <= -1 and q value< 0.05 were considered to differentially expressed lincRNAs or PCGs.

**2.3 Co-expression network and WGCNA analysis**

An expression matrix including PCGs and lincRNAs across all datasets was used to construct a co-expression network in *Arabidopsis thaliana*. In brief, Pearson correlation coefficients were calculated between all lincRNAs and PCGs. Then the correlation value was transformed by Fisher transformation, and transformed fisher values were standard normalized to obtain z-scores whose co-expression edge value distribution has a mean of zero and a standard deviation of one (standard normal). Finally, a cutoff z-score 2.0 was used to obtain significant edges between PCGs and lincRNAs.

The co-expression network of lincRNAs and PCGs was constructed individually for *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Capsella rubella*, *Brassica napus*, *Marchantia polymorpha*, *Oryza sativa*, and *Zea mays* using WGCNA (Langfelder and Horvath 2008). First, PCGs and lincRNAs with a low coefficient of variation (CV<0.7) among samples were filtered out (Tian et al. 2019). The expression level (TPM) of lincRNAs and PCGs was then log2 transformed and normalized into z-score. The soft power of 12 was used for fitting the scale-free topology of the co-expression network. The default parameters of the dynamic tree were used to get modules of the co-expression networks. The eigengenes of the modules were computed from the first component of the module expression matrix. The PCGs within modules were set as the input of GO enrichment analysis by GOseq (Young et al. 2010). Visualization of the co-expression network was done by Cytoscape (Shannon et al. 2003).

LncRNAs regulate their target genes in *cis* or/and *trans*. To identify potential *cis* targets of lincRNAs, upstream and downstream PCGs which are in close physical proximity to lincRNAs in the genome were considered. For the prediction of *trans* regulation, protein coding genes that co-expressed with lincRNAs in the network above are considered to be *trans* targets of lincRNAs.

**2.4 DNase-seq, ChIP-seq analysis, and identification of enhancers**

ChIP-seq data for AP1 and SEP3 as well as DNase-seq data were obtained from (Pajoro et al. 2014; Chen et al. 2019; **Supplemental Table S13**). Regarding the histone modification ChIP-seq and GRO-seq data, the raw reads were downloaded from the NCBI SRA database (**Supplemental Table S13**). The read adapters were trimmed and filtered, followed by mapping to the *Arabidopsis* genome using the hisat2 software. Uniquely mapped reads and merged replicates were used for MACS2 (Zhang et al. 2008) software with default parameters to obtain significant TF-bound genomic regions ('peaks') or DNase I hypersensitive sites (DHSs) (Chen et al. 2018). Bigwig files were created by MACS2. Heatmap and line plots of coverage

maps were generated using deepTools2 (Ramírez et al. 2016).

Identification of enhancers in *Arabidopsis* followed the definition as in (Zhu et al. 2015b). Briefly, intergenic DHS regions with a distance of more than 1.5 kb from the closest TSS are considered to be enhancers in *Arabidopsis*. The final set of enhancers in this study were obtained from merging results from DNase-seq data at days 0, 2, 4, 8 after induction, and also published enhancers identified in (Zhu et al. 2015b).

**2.5 Gene ontology (GO) enrichment analysis**

To functionally annotate sets of PCGs, GOseq (Young et al. 2010) and clusterProfiler (https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) were used for gene ontology (GO) enrichment analysis with default parameters. All detected genes in the whole *Arabidopsis* gene catalog were used as the background. GO annotations were obtained from TAIR10 (https://www.arabidopsis.org/).

**2.6 Tissue specificity index**

The Tissue specificity index (TSI) was determined to measure the expression specificity of PCGs and lincRNAs using the input expression matrix calculated from RNA-seq datasets (**Supplemental table S1**) (Kryuchkova-mostacci and Robinson-rechavi 2017).

Xi is the expression of the gene in tissue i while n is the number of tissues. The higher TSI is, the more specific the locus is. TSI is defined as:

$$TSI = \frac{\max\limits_{1 \le i \le n} (x_i)}{\sum_{i=1}^{n} x_i}.$$

**2.7 Quantitative Reverse Transcriptase-PCR (qRT-PCR)**

For validation of lincRNAs expression level in *Arabidopsis* flower tissues, total RNA was extracted from each developmental time point in three biological replicates with the GenUP™ Plant RNA Kit (#350701502 from Biozym Scientific GmbH) including an on-column DNaseI digest. The cDNA was synthesized by the kit ProtoScript® II First Strand cDNA Synthesis Kit (#E6560S from NEB). The *TIP41* was used as the reference gene. A primer list is provided in **Supplemental Table S12**. The relative expression of lincRNAs was calculated according to a standard method (Livak and Schmittgen 2001).

**2.8 Identification of lincRNA family, homologous lincRNAs and lincRNA evolutionary age group**

The repeats masked lincRNA sequences from each plant species were reciprocally compared with each other by BLAST 2.4.0+ (-evalue 1e-5 -num_threads 10 -max_target_seqs

1 -word_size 8 -strand plus -outfmt 6). LincRNA sequences of two plant species with an alignment E-value < 1e-5 were considered to be the best hits and were considered to be homologs (Hezroni et al. 2015). To identify lincRNA family, a lincRNA sequence similarity network was built to connect homologous lincRNAs from each species. An unsupervised graph cluster algorithm (MCL, https://micans.org/mcl/) was then used to identify lincRNA cluster within the constructed network (with the parameter: --abc -I 2.0). Each cluster of homologous lincRNAs was designated a lincRNA family that was then assigned to one of the three types of families: one2one, one2many and many2many, based on the number of homologous lincRNAs in the plant species from which the lincRNA(s) were identified. If a lincRNA has only a single homolog in all plant species with the homologous linRNA identified, the cluster contains these homologous lincRNAs was defined as a one2one family; if a lincRNA has multiple homologs (≥2) in at least one of the plant species, the cluster contains the homologous lincRNAs was defined as a one2many family; if a lincRNA has multiple homologs (≥2) in all plant species with the homologous lincRNAs, the cluster contains the homologous lincRNAs was defined as a many2many family.

The MCScanX (Wang et al. 2012) software was used to identify syntenic regions between two species based on pairwise comparisons. PCGs within the syntenic regions were used to define syntenic (conserved) lincRNAs between the two corresponding species. We considered three PGCs at each side of a given lincRNA. A lincRNA that was found in two plant species, flanked by a minimum of one syntenic PCG on each side and had a minimum of three syntenic PCGs was defined as syntenic lincRNA (Pegueroles et al. 2019).

A set of different criteria were used to identify different evolutionary age groups of conserved lincRNAs across different levels of plant lineages, including Plants, Angiosperms, Monocots, Eudicots, and Brassicaceae. LincRNAs conserved in the evolutionary age group of Plants should have a homolog in *Amborella trichopoda*, at least a homolog in one of the eudicots, one of the monocots, and one of non-flowering plants (i.e. Atr& (Ath|Aly|Cru|Bol|Bna|Bra|Bju|Aal|Cla|Csa|Car|Gma|Fve|Vvi|Sly|Slo) & (Osa|Zma) & (Ppa|Mpo|Cre|Vca|Smo|Afi)). LincRNAs conserved in the evolutionary age group of Angiosperms should have homolog in *Amborella trichopoda*, at least one homolog in eudicots and one homolog in monocots (i.e. Atr& (Ath|Aly|Cru|Bol|Bna|Bra|Bju|Aal|Cla|Csa|Car|Gma|Fve|Vvi|Sly|Slo) & (Osa|Zma)). LincRNAs conserved in monocots and eudicots the evolutionary age group of Moncots_Eudicots (i.e both monocots and eudicots) should have at least one homolog in both monocots and eudicots (i.e.

(Ath|Aly|Cru|Bol|Bna|Bra|Bju|Aal|Cla|Csa|Car|Gma|Fve|Vvi|Sly|Slo) & (Osa|Zma)). LincRNAs conserved in the evolutionary age group of Eudicots should have a homolog in Sacred Lotus (*Nelumbo nucifera*, a basal eudicot), at least one homolog in other eudicots (i.e. Slo & (Ath|Aly|Cru|Bol|Bna|Bra|Bju|Aal|Cla|Csa|Car|Gma|Fve|Vvi|Sly)). LincRNAs conserved in the evolutionary age group of Monocots should have homolog in both *Oryza sativa* and *Zea mays* (Osa|Zma).

LincRNAs conserved in the evolutionary age group of Brassicaceae should have a homolog in at least two species of *Brassicaceae* and also have at least one homolog in *Brassicaceae* lineage I (Ath, Aly, Cru) and II (Bol, Bra, Bna, Bju, Aal), respectively.

Old lincRNAs were defined as those found in the evolutionary age groups of Plants, Angiosperms, and Eudicots while young lincRNAs were defined as those found in the evolutionary age group of Brassicaceae.

**2.9 Identification of peaks of histone modification and TF binding overlapping with lincRNA and PCGs**

Peak files of histone modifications and TFs were obtained from the ChIP-Hub database (Chen et al. 2019) in *Arabidopsis thaliana*. The peaks overlapping with the 1kb upstream/downstream regions of lincRNAs and PCGs were retrieved by the intersect function of the bedtools v2.25.0. The frequency of lncRNAs or PCGs with histone modification or TF-binding site was calculated by the number of overlapping sites/the total number of lincRNAs or PCGs.

**2.10 Identification of transposable elements (TEs) overlapping with lincRNAs**

TEs from *Arabidopsis thaliana* were downloaded from TAIR10 (https://www.arabidopsis.org/download_files/Genes/TAIR10_genome_release/TAIR10_transposable_elements/TAIR10_Transposable_Elements.txt). TEs in other plant genomes were identified by EDTA (Ou et al. 2019, https://github.com/oushujun/EDTA). The parameters for plant genomes other than rice and maize are EDTA.pl --genome genome.fasta --species others --cds cds.fa --anno 1 --threads 20. For genomes of rice and maize, the “--species” parameters were set Rice or Maize, respectively. The function of intersect of the bedtools v2.25.0 was used to identify lincRNAs and PCGs intersecting with TEs using the criterion of at least 1 nt overlapping.

# 3. Results

## 3.1 Transcription factor-mediated activities of enhancer lincRNAs in flower development

### 3.1.1 Genome wide identification of flower-related lincRNAs in Arabidopsis

LincRNAs are defined as >200 nt transcripts without protein coding potential in the intergenic regions. In order to globally identify lincRNAs involved in *Arabidopsis* flowering and floral organ development, we used both polyA and total RNA-seq datasets from an inducible system for synchronized flower induction based on pAP1:AP1-GR *ap1 cal* line (chemically inducible AP1-GR protein under the pAP1 promoter in *ap1 cal* double mutant background) (Chen et al. 2018). We studied lincRNA expression levels focusing on stage 0 (uninduced, inflorescence meristem), stage ~2 (2 days after induction (DAI), meristem specification), stage ~5 (4 DAI, floral whorl specification), stage ~8/9 (8 DAI, floral organ differentiation) (**Figure 3.1A**). Additionally, we also collected RNA-seq datasets from diverse developmental stages and tissues in public databases (e.g. NCBI SRA), with a major focus on reproductive meristematic stages and floral organ development (**Supplemental Table S1**). With the bioinformatics pipeline described above (details described in **Methods**, **Figure 3.1B**), we obtained a set of 4106 flower related lincRNAs that show transcriptional activity in reproductive meristems or flowers (**Supplemental table S2**). Compared to 6480 published lincRNAs (Liu et al. 2012), ~9.5% lncRNAs in our flower-related dataset are shared (**Figure S1A**) which is suggestive of enhancing coverage and diversity of lincRNAs in flower development. Additionally, we collected 19740 published lincRNAs from other studies (Liu et al. 2012; Di et al. 2014; Zhu et al. 2014; Wang et al. 2014; Okamoto et al. 2010; Matsui et al. 2008; Szcześniak et al. 2016) and 37.5% of them are common with 4106 flower related lincRNAs (**Figure S1A**). This implies that current lincRNA datasets in *Arabidopsis* flowers are still incomplete, possibly because of the high expression specificity and low overall abundance. To investigate differences and similarities of lincRNAs and PCGs (protein-coding genes), we compared transcript length, evolutionary conservation, and expression pattern. We found that lincRNAs are on average shorter than PCGs (**Figure S1B**), and comparable with published ones, supporting the quality of our datasets. The median length of lincRNAs is 438 bp (mean: 762 bp) while the median length of PCGs is 1909 bp (mean: 2206 bp). The evolutionary sequence conservation of lincRNAs, as estimated by PhyloP scores (Haudry et al. 2013), is lower than that of PCG exons, but higher than that of introns (**Figure S1C**), in agreement with an earlier study (Yuan et al. 2016). LincRNAs overall tend to be more lowly expressed than PCGs (**Figure S1D**) and show higher levels of tissue-specificity (**Figure S1E**).

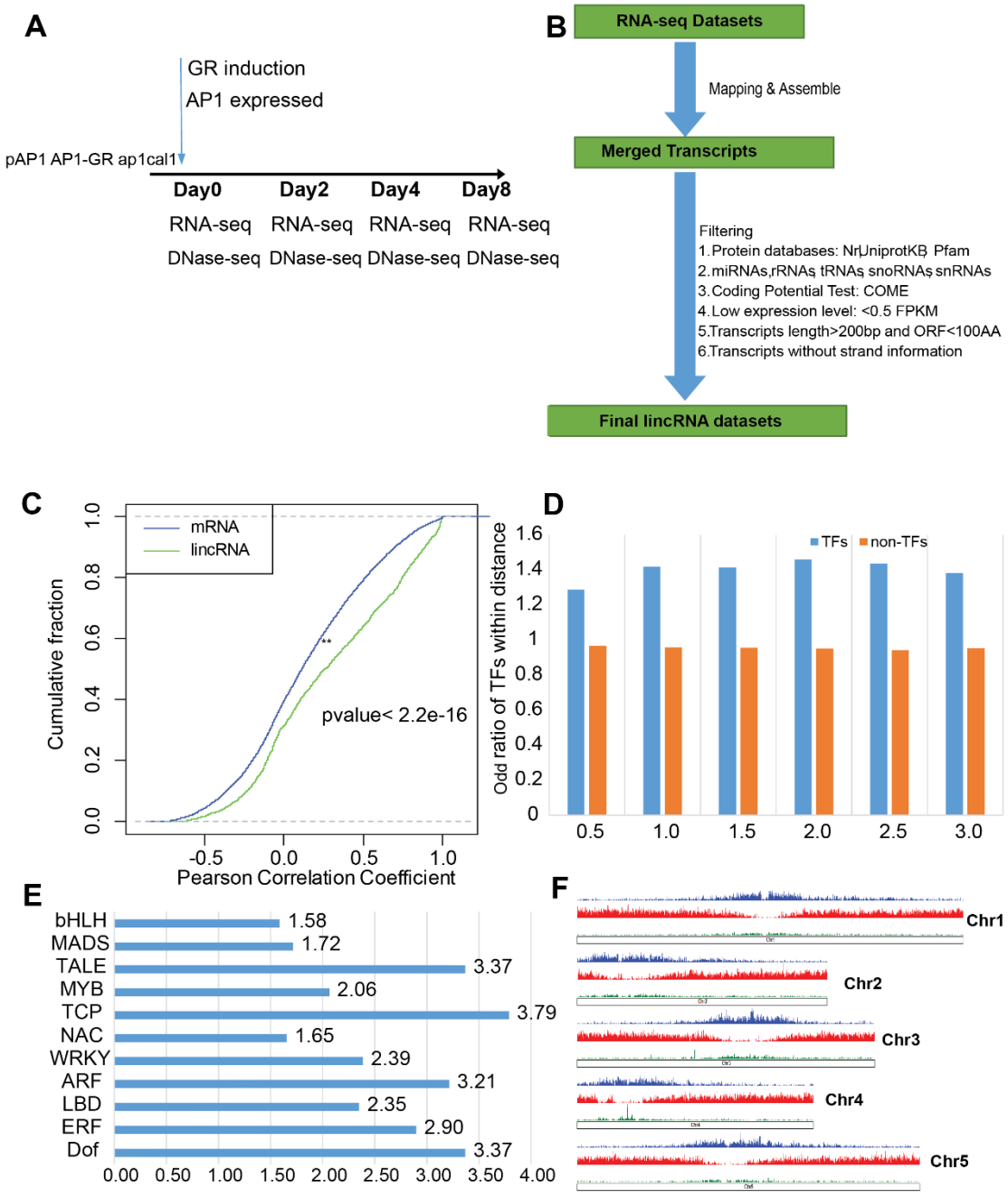



**Figure 3.1**: **Spatial distribution of lincRNAs in the *Arabidopsis* genome.** (A) Our experimental setup for the identification of lincRNAs in flower development. (B) Overview of the computational pipeline for lincRNA identification (details described in methods). (C) Pearson expression correlation of lincRNAs and all expressed PCGs with its nearest neighbor genes, p-value calculated by Wilcoxon rank-sum test. (D) Extent of PCG enrichment within different distance ranges to lincRNA loci (0.5 kb, 1.0 kb, 1.5 kb, 2.0 kb, 2.5 kb and 3.0 kb). TF: Transcription factor; non-TF: PCGs which are not TFs. (E) Certain TF gene families, such as MADS and TCP TFs, are enriched as the neighboring PCGs of lincRNAs determined by Fisher exact test (p-value <0.05, Odds ratio of TF = (number of overlap/number of specified neighboring PCGs)/(number of TFs/ number of total PCGs). (F) The distribution of TEs (blue),

PCGs (red), and lincRNAs (green) in *Arabidopsis* chromosomes (Chr1, Chr2, Chr3, Chr4, and Chr5).

The genome position of lincRNAs can provide clues about the functions and mechanisms of lincRNAs. LincRNAs are often associated with their neighboring PCGs (Kopp and Mendell 2018). In agreement with this and with previous results (Luo et al. 2016), we found that lincRNAs are more co-expressed with the nearest PCG compared to neighboring PCG-PCG pairs (p value<2.2e-16) (**Figure 3.1C**). This indicates that lincRNAs are potentially co-regulated with neighboring PCGs in *cis*. Additionally, these neighboring PCGs are preferentially enriched in members of certain transcription factor (TF) family genes when compared to a control set of PCGs (odds of enrichment: 1.74 vs 1, determined by Fisher exact test). In order to confirm this, two methods were utilized. Firstly, different maximum distances (0.5 kb, 1.0 kb, 1.5 kb, 2.0 kb, 2.5 kb, and 3.0 kb) were used to co-locate lincRNAs and PCGs, and the extent of TF gene enrichment within this range is very similar (**Figure 3.1D**). Secondly, we used GenometriCorr (Favorov et al. 2012) to analyze this trend. We found that compared to features like signaling and structural genes, TF genes are more significantly physically associated with lincRNAs loci. Furthermore, we found that MADS and TCP TF gene families are enriched as neighboring genes to lincRNAs (**Figure 3.1E**). Additionally, lincRNAs are distributed across the chromosome and reveal a preference towards pericentromeric regions, similar to TEs but distinct from PCGs (1431 lincRNAs, 69.6%) (**Figure 3.1F**). These pericentromeric lincRNAs are apparently highly expressed in actively dividing tissues (e.g. anthers and meristems) (**Supplemental Table S3**). Additionally, we found that the most abundant TE types associated with pericentromeric lincRNAs are RC/Helitron and LTR/Gypsy (**Figure S2A; Supplemental Table S4**), which is consistent with high enrichment of pericentromeric histone CENH3 (**Figure S2B, S2C**). In summary, genome wide identification of lincRNAs in flower development reveals the non-random distribution of lincRNAs in the *Arabidopsis* genome.

### 3.1.2 Flower-related lincRNAs display associated expression with different regulatory modules

We studied lincRNA expression levels focusing on stage 0 (uninduced, inflorescence meristem), stage ~2 (2 days after induction (DAI), meristem specification), stage ~5 (4 DAI, floral whorl specification), stage ~9 (8 DAI, floral organ differentiation) with the floral induction system. Furthermore, other public datasets obtained from flower tissues were also collected for the construction of a co-expression network (**Supplemental Table S1**), which allowed us to investigate the dynamic expression of lincRNAs and PCGs during floral development processes. In total, 337 lincRNAs are differentially expressed (FDR<0.05, $|\log_2(FC)|>=1$) in the flower developmental time-series (0, 2, 4, 8 DAI) and other relevant

comparisons (**Supplemental Table S5**). To confirm the expression pattern of expressed lincRNAs in flower development, 12 expressed or moderately expressed lincRNAs in floral tissues were chosen for reverse transcription quantitative real-time PCR (RT-qPCR) verification. The qRT-PCR experiments were performed by Johanna Müschner. We found that the lincRNAs show no differences between qPCR and RNA-seq values at early developmental stages (day 0, 2, 4 after AP1 induction) and the expression pattern is consistent. The results show that RNA-seq data and qPCR results are largely consistent (**Figure 3.4**). Additionally, the percentage of differentially expressed lincRNAs and PCGs are overall consistent and demonstrate prevalent changes at later stages (e.g. 4, 8 DAI) of flower development (**Figure S3A)**. To study the association of lincRNAs and their neighboring PCGs, we compared changes in expression across developmental stages. We found a clear prevalence for concerted up- and down-regulation of lincRNAs and their nearest PCG neighbor (**Figure S3B, Supplemental table S6**), suggesting a common control mechanism. Furthermore, lincRNAs and the nearest neighboring target PCGs are often simultaneously marked by H3K4me3 or H3K27me3 (**Figure S3C**, determined by Fisher's exact test, see details in **Supplemental Table S7**), and this trend holds across tissues and stages (**Supplemental table S7**).

To investigate the potential roles of lincRNAs, we constructed a lincRNA-PCGs co-expression network by WGCNA (Langfelder and Horvath 2008) with the soft threshold power beta=12 (**Figure S4A**) which makes use of correlations among genes and cluster these genes into modules. The co-expression network was divided into 62 modules according to the expression pattern in each module (**Supplemental table S8**). All PCGs in each module of the network were submitted to GO enrichment analysis (**Supplemental Table S9**). Eight flower-related modules (# 5, 9, 13, 15, 35, 45, 46, and 58) were chosen and significantly enriched GO categories were summarized in **Figure 3.2A**. For example, GO analysis of module 58 suggests functions in vegetative to the reproductive phase transition, and in meristem structural organization, while genes in Module 5 module preferably act in pollen development (**Figure 3.2A**). Expression patterns (eigenvalues) were also analyzed for each module (**Figure 3.2B**), which is essentially consistent with the functional annotation of each module. For instance, genes in module 58 have preferred roles of vegetative to the reproductive phase transition of the meristem, whereas it is highly expressed in flowering transition meristem stages (**Figure S5**) (Klepikova et al. 2015). Genes in module 35 have predominant roles in the maintenance of shoot apical meristem identity according to GO enrichment analysis (p value= 0.018008), which is consistent with the expression pattern of meristem marker *SHOOTMERISTEMLESS* (*STM*) (**Figure S5**). Furthermore, lincRNAs in modules 58 and 35 are co-expressed with many

flower related genes (**Supplemental Table S10, S11**). For example, AtklncRNA3562 in module 58 (vegetative to the reproductive phase transition of meristem and meristem structural organization) is mostly expressed in meristems at floral transition and in carpels (**Figure 3.3C**) and shows concerted expression with flower-related genes including *REVOLUTA* (*REV*) and *VARIANT IN METHYLATION 3* (*VIM3*) within the same module (**Figure 3.3A, B; Supplemental table S10**). AtklncRNA15638 in module 35 (maintenance of the shoot apical meristem identity) is co-expressed with *AINTEGUMENTA (ANT)* and *AGAMOUS-LIKE 24 (AGL24)* (**Figure 3.3D, E; Supplemental table S11**). It is highly expressed in the shoot apex at floral transition (e.g. M5, **Figure 3.3F**) and correlated with the repression of *FLC* (**Figure S5**). Taken together, flower-related lincRNAs display associated expression with different regulatory modules.

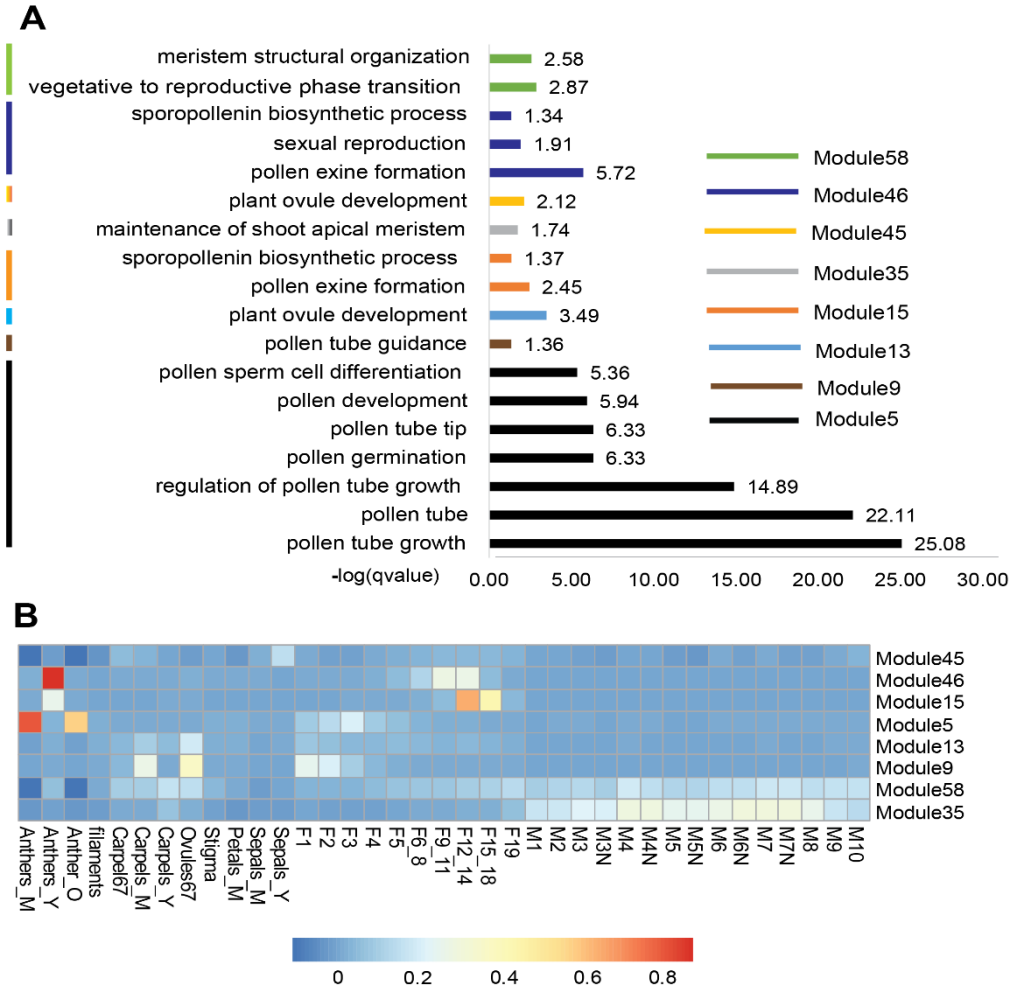



**Figure 3.2**: **Prediction of lincRNA functions using a lincRNA-PCG co-expression network** (Guilt-by-Association). (**A**) GO functional annotation of eight flower related module (# 5, 9, 13, 15, 35, 45, 46, and 58). (**B**) The expression pattern of each module with eigenvalues (1st principal component) is shown by the heatmap in diverse floral and meristem tissues.

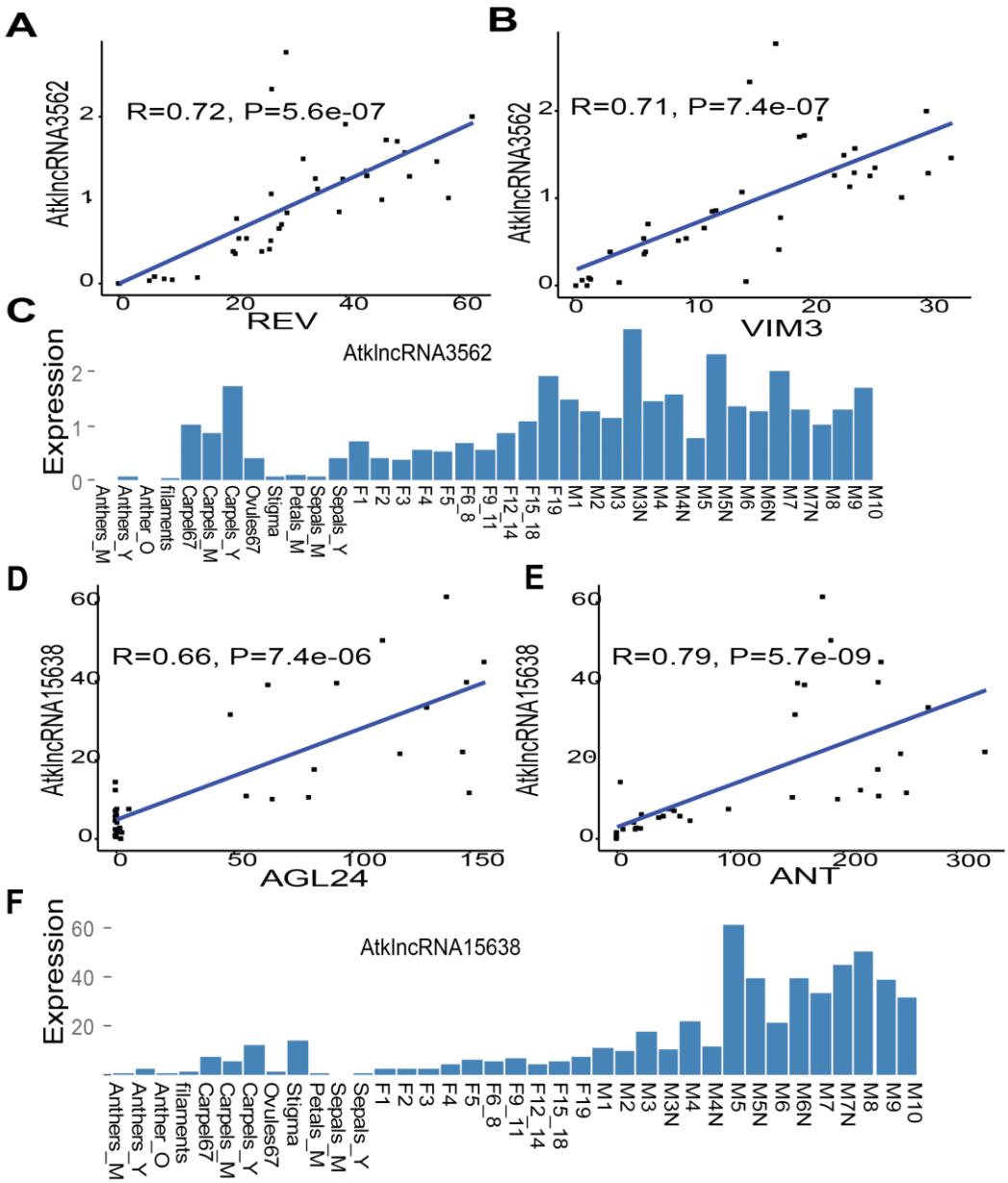



**Figure 3.3**: **Prediction of functions of lincRNAs using a lincRNA-PCG co-expression network** (Guilt-by-Association). (A) Scatter plot of expression for AtklncRNA3562 and *REV* (*REVOLUTA*). (B) Scatter plot of expression for AtklncRNA3562 and *VIM3* (*VARIANT IN METHYLATION 3*). (C) Expression pattern of AtklncRNA3562 across diverse tissues. (D) Scatter plot of expression for AtklncRNA15638 and *AGL24 (AGAMOUS-LIKE 24)*. (E) Scatter plot of expression for AtklncRNA15638 and *ANT* (*AINTEGUMENTA*). (F) Expression pattern of AtklncRNA15638 across diverse tissues.

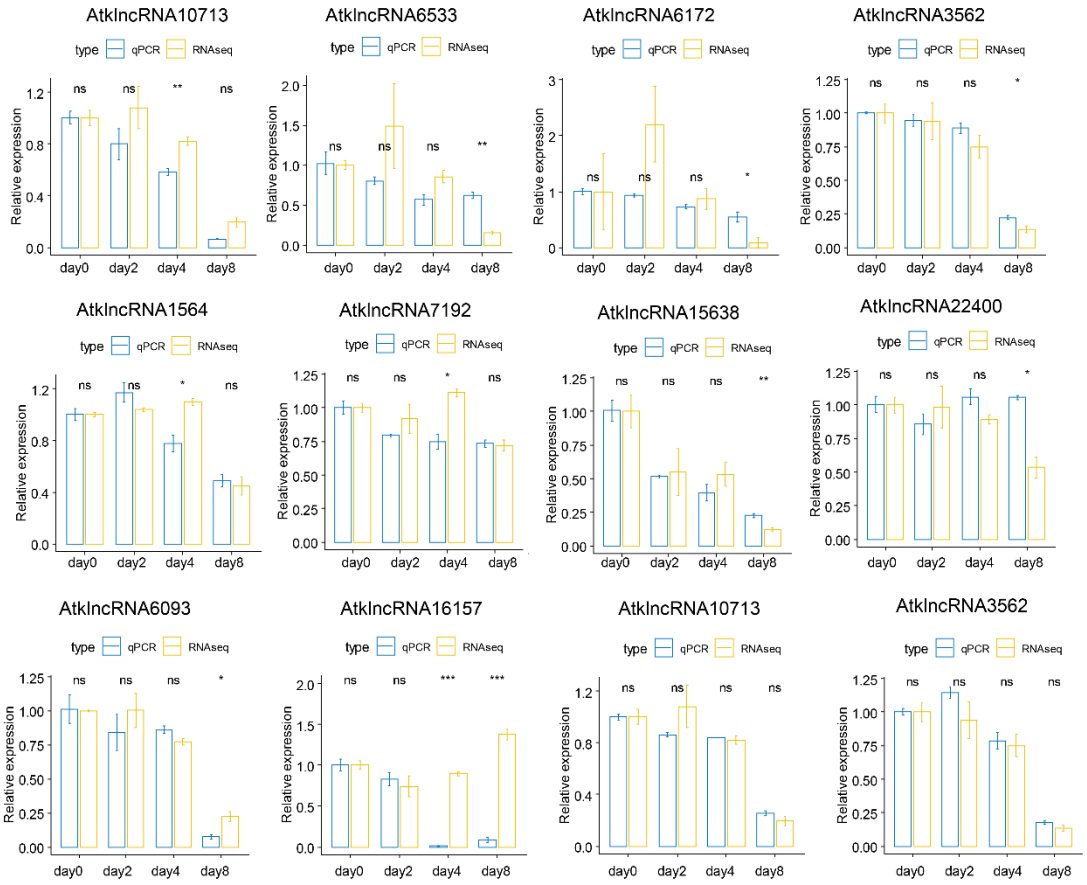



**Figure 3.4**: **RT-qPCR validation of 12 lincRNAs identified by RNA-seq**. The expression of day0 used as reference for calculation of p values. *, P value <0.05; **, P value <0.01; ***, P value <0.001; ns, not significant.

### 3.1.3 Flower-related lincRNAs are enriched in genomic regions bound by developmental master TFs and in enhancers

LincRNAs that are transcribed from enhancers by Pol II activity have been defined as enhancer RNAs (eRNAs) (Shlyueva et al. 2014). Flower development is regulated by key regulatory TFs including floral homeotic and floral meristem identity factors (e.g. AP1, AP2, AP3, PI, AG, SEP3, LFY) (Krizek and Fletcher 2005). These master regulatory TFs are enriched in enhancers during flower development (Yan et al. 2019). Potential direct target genes of these factors have been identified by ChIP-seq in combination with genome-wide expression analyses; however, most regulatory interactions have not yet been studied in detail. In order to understand the relationship between enhancers and lincRNAs, distal regions (1.5kb distance from TSS of the gene) with open chromatin/accessibility are marked enhancers (Zhu et al. 2015b). These enhancer regions demonstrate high enrichment of open chromatin (**Figure 3.5A**) and high Pol II occupancy suggesting transcriptional activity of enhancers (**Figure 3.5B**). Additionally, these enhancers are bi-directionally transcribed, as indicated by GRO-seq

data (**Figure 3.5C**), polyA RNA-seq (**Figure 3.5D**), and total RNA-seq (**Figure 3.5E**). This is consistent with identified putative enhancer-like elements (PEs) in *Arabidopsis* (Wang and Chekanova 2019). We repeatedly sampled 1000 genomic regions not overlapped the lincRNA region. Then we overlapped these random regions with enhancer regions and computed enrichment values (overlap_num/shuffle_overlap_num). Random regions sampled from the genome have no enrichment (nearly 1) with enhancer regions. Interestingly, we find that lincRNAs are often associated with these enhancers (**Figure 3.6A**), which are validated in recent studies of enhancers in *Arabidopsis thaliana* (Yan et al. 2019; Zhu et al. 2015b). We define these lincRNA-associated enhancers as enhancers whose distance between lincRNAs and enhancers should be less than 300 bp (Gil et al. 2018). Unexpectedly, these lincRNA-associated enhancers (la-e) display higher chromatin accessibility (**Figure 3.6B**) and Pol II occupancy (**Figure 3.6C**) compared to non-lincRNAs associated enhancers (na-e). LincRNA transcripts are there, which might be associated with higher enrichment of open chromatin. It suggests that lincRNAs expression level or just lincRNA transcription contributes to chromatin open. Data from the human field reveals eRNA drives transcription by promoting chromatin accessibility (Melo et al. 2013; Mousavi et al. 2013; Azofeifa et al. 2018). We next asked whether floral master regulatory TFs preferentially bind to lincRNA-associated enhancers. Indeed, there is a strong enrichment in TF binding (**Figure 3.6D, E**; **Figure 3.7A**). For example, we found that 92% (138/150) of the AP1-bound lincRNA loci were characterized by an open chromatin state, and lincRNAs are significantly associated with AP1 binding (**Figure 3.7A, B**), which are suggestive that lincRNA activity may be the consequence of AP1 binding. Most of the TF-bound lincRNA loci (e.g. 56% of 138 AP1-binding associated lincRNAs) are associated with distal open chromatin sites that are typically defined as enhancers in the genome (**Figure 3.7A, B, C**). This indicates that these lincRNAs could be classified as enhancer-associated RNAs. Enhancers bound by AP1 or SEP3 are significantly (p value<0.05) associated with active lincRNAs. For example, 18.2% of AP1-bound enhancers are significantly (p value=0.0006) linked with detectable lincRNAs expression (**Supplemental Table S14**). Interestingly, TF binding sites are most strongly enriched within lincRNA gene bodies (**Figure 3.7B, C**). This is different from typical PCGs, where TF binding sites are mostly enriched in promoters and downstream regions of the genes. This is possibly because of lincRNAs overlapping with enhancers which are bound with multiple master floral TFs such as AP1 and SEP3. In agreement with their residence in open chromatin regions, histone marks are usually depleted from lincRNA loci, particularly from their gene bodies (**Figure 3.7B, C**) in contrast to TF-regulated PCGs. In summary, most lincRNAs display patterns of TF binding and histone

modification that are very different from PCGs and flower-related lincRNAs are associated with master TFs and enhancers.

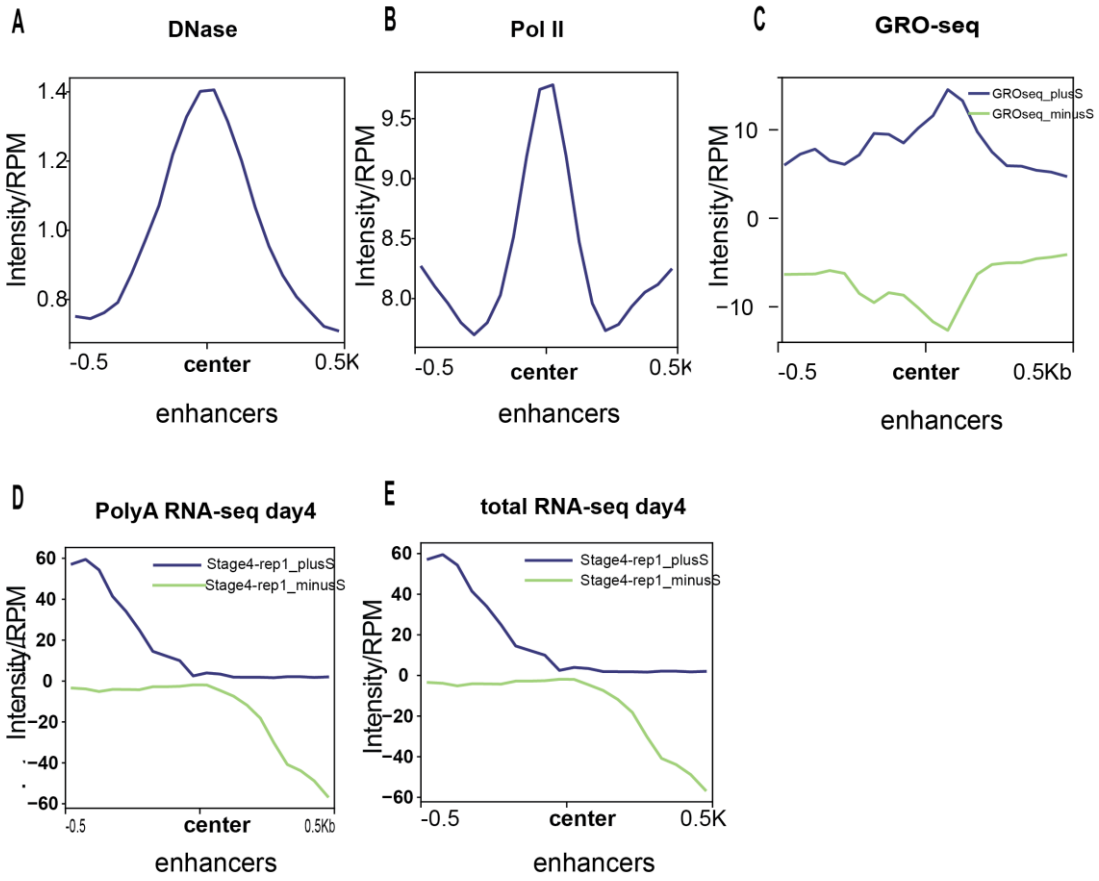



**Figure 3.5: Chromatin marks linked with active transcription in enhancers**. (A) DNase (open chromatin/accessibility) enrichment signal for identified enhancers. (B) PolII occupancy signal for identified enhancers. (C) Signal of GRO-seq in flower buds for newly identified enhancers. (D) Signal of the polyA RNA-seq data 4 DAI for identified enhancers. (E) Signal of the total RNA-seq data 4 DAI for identified enhancers.

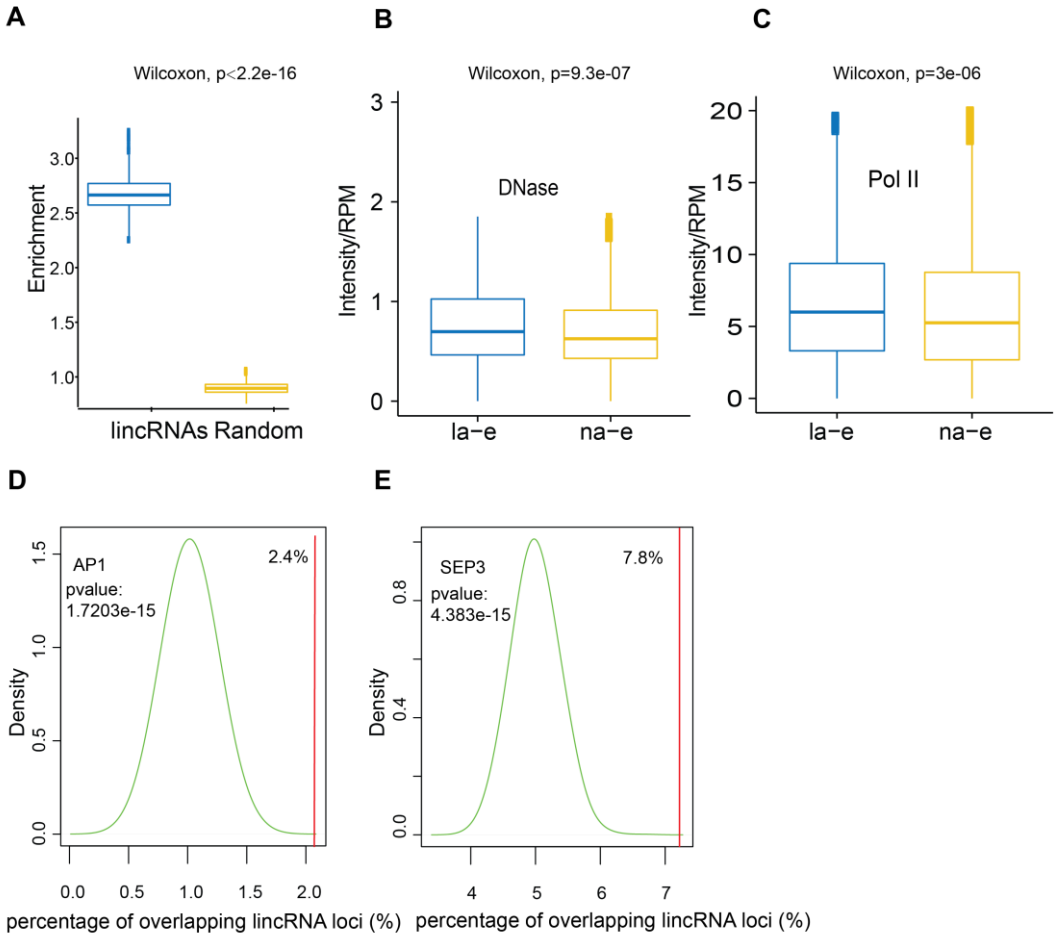



**Figure 3.6: Flower related lincRNAs are associated with master TF regulators and enhancers**. (A) LincRNAs are associated with enhancers. We repeatedly sampled 1000 genomic regions not overlapped the lincRNA region. Then we overlapped these random regions with enhancer regions and computed enrichment values (overlap_num/shuffle_overlap_num). Random regions sampled from the genome have no enrichment (nearly 1) with enhancer regions. (B) lincRNAs associated enhancers (la-e) have higher accessibility signal than that of non-lincRNAs associated enhancers (na-e). (C) lincRNAs associated enhancers (la-e) have a higher Pol II transcriptional signal than that of non-lincRNAs associated enhancers (na-e). (D) Flower related lincRNAs are associated with AP1 binding. Green lines, distribution of the percentage of lincRNAs overlapping with a random sampling of AP1 ChIP-seq peaks. Red lines, the actual percentage of lincRNAs loci overlapping with AP1 ChIP-seq peaks. (E) Flower related lincRNAs are associated with SEP3 binding. Green lines, distribution of the percentage of lincRNAs overlapping with a random sampling of SEP3 ChIP-seq peaks. Red lines, the actual percentage of lincRNAs loci overlapping with SEP3 ChIP-seq peaks.

A

| | AG | AP1 | AP2 | AP3 | LFY | PI | SEP3 |
|---|---|---|---|---|---|---|---|
| Target lincRNA | 197 | 150 | 157 | 198 | 402 | 218 | 475 |
| Open Chromatin | 151(77%) | 138(92%) | 130(83%) | 172(87%) | 305(76%) | 194(89%) | 408(86%) |
| enhancer lincRNAs | 87(58%) | 77(56%) | 73(56%) | 94(55%) | 175(57%) | 101(52%) | 230(56%) |

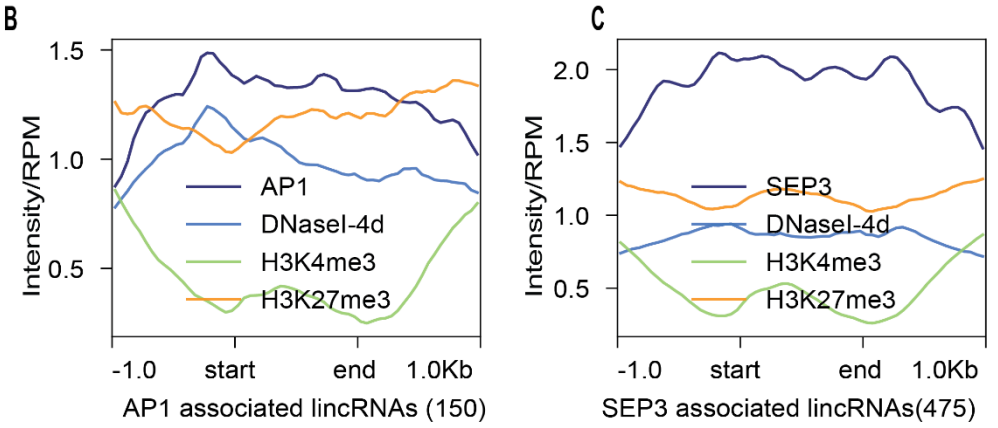



**Figure 3.7: Flower related lincRNAs are associated with master TFs**. (A) Candidate lincRNAs that are associated with DNA-binding by master regulatory TFs. (B) Signal of AP1 ChIP-seq, DNaseI-seq 4d, H3K4me3, and H3K27me3 for 150 AP1 regulated candidate lincRNAs. (C) Signal of SEP3 ChIP-seq, DNaseI-seq 4d, H3K4me3, and H3K27me3 for 475 SEP3 regulated candidate lincRNAs.

### 3.1.4 Chromatin states of flower related lincRNAs that are active in flowers

Previously, the chromatin landscape of *Arabidopsis thaliana* has been classified into 9 different states based on patterns of histone methylation, CG methylation, GC content, and histone variants (Sequeira-Mendes et al. 2014). In order to systematically study the chromatin environment of flower related lincRNAs, we used ChromHMM (Ernst and Kellis 2017) to infer chromatin states in floral tissues with collected histone modification and DNase-seq datasets from shoot apical meristem and floral tissues (**Figure 3.8A**). In order to differentiate TEs on lincRNAs, we classify lincRNAs into lincRNA overlapping with TEs (TE-lincRNAs) and lincRNA not overlapping with TEs (non-TE-lincRNAs). The chromatin state of TEs is different from that of TE-associated lincRNAs and non-TE associated lincRNAs (**Figure 3.8B**). Additionally, the chromatin state in flowers reflects the differences of chromatin states of lincRNAs: TE-lincRNAs and non-TE-lincRNAs. TE-lincRNAs are enriched with flowerCS5, 6,7,8,9 (flowerCS5: closed chromatin, H3K9me2, TE-enrich; flowerCS6: TE-enrich; flowerCS7: open chromatin, enhancer-enrich; flowerCS8/9: H3K27me3-enrich) in flowers. Moreover, they are associated

with two kinds of heterchromatin states: constitutive (H3K9me2, AGDP1) and development-related heterochromatin (H3K27me3). For example, the lincRNA Aralnc.19175/*linc-AP2* are overlapping with both TEs and enhancers and thus it resides in flowerCS6 and flowerCS7 (**Figure 3.8C**). In summary, flower related lincRNAs are associated with distinct chromatin states in the *Arabidopsis* genome.

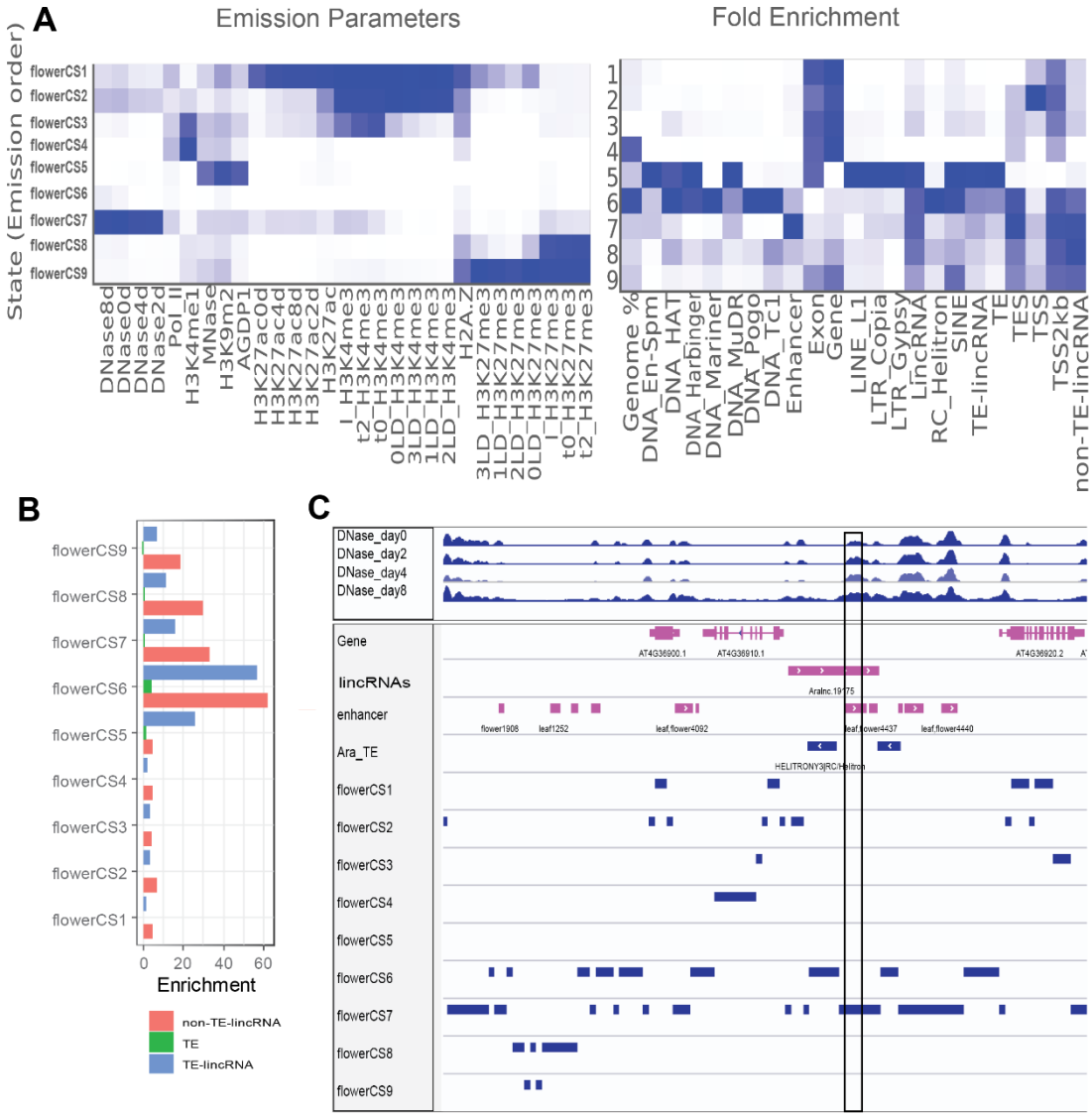



**Figure 3.8**: **LincRNAs are associated with diverse chromatin states in flowers**. (A) ChromHMM (Ernst and Kellis 2017) was used to infer chromatin states in floral tissues with collected histone modification (H3K4me1, H3K27ac, H3K4me3, and H3K27me3) and DNase-seq datasets from shoot apical meristem and floral tissues. (B) Different chromatin states in flowers are associated with TE-associated lincRNAs (TE-lincRNAs), non-TE-associated lincRNA (non-TE-lincRNAs), and TEs. (C) Chromatin states for Aralnc.19175/*linc-AP2* in flowers (flowerCS5: closed chromatin, H3K9me2, TE-enrich; flowerCS6: TE-enrich; flowerCS7: open chromatin, enhancer-like.; flowerCS8/9: H3K27me3-enrich). Gene: protein coding genes; Ara_TE: TEs in *Arabidopsis*; enhancer: enhancers.

### 3.1.5 Flower related lincRNAs are associated with floral gene regulation

To better understand the dynamic relationship of TF binding, lincRNA accumulation, and chromatin status, we made use of ChIP-seq, RNA-seq, and DNase-seq time-series datasets that have been generated in the AP1-based system for synchronized floral induction (Pajoro et al. 2014; Yan et al. 2019). We confirmed high positive Pearson correlation coefficients values between change in AP1 binding from day 2 to day 4 (AP1.day2_4, the same time point) and change in chromatin accessibility from the later time points (DNase.day4_8, **Figure S6A**) (Pajoro et al. 2014). The same was found for SEP3 (**Figure S6B**). These findings re-confirm that AP1 and SEP3 binding precedes changes in chromatin accessibility, suggesting roles as pioneer factors (Pajoro et al. 2014). Now, we aimed at addressing whether lincRNA activity precedes enhancer opening, or is a consequence of this process, possibly as a by-product of its activity. The fold change in AP1 binding from day 2 to day 4 and fold expression change of enhancer-associated lincRNAs from the same time point are significantly correlated (**Figure 3.9A**), which indicates that AP1 binding could induce transcription of enhancer-associated lincRNAs. A similar result was obtained for SEP3 (**Figure 3.9B**). However, the majority of the lincRNAs (75.3%, 113/150) do not display any change of expression while they are bound by AP1. Additionally, the majority of SEP3-bound lincRNAs (73.0%, 347/475) do not display any change of expression, either. Therefore, the role of AP1/SEP3 in the activation of lincRNAs is only limited to a subset of them. The reason behind this is that the majority of SEP3/AP1-bound lincRNAs might be unstable RNAs that cannot be detected through general polyA or total RNA-seq. Furthermore, these unstable RNAs can only be seen in nascent RNA-seq such as GRO-seq and exosome mutants (because they can be degraded by exosome complex quickly) (Thieffry et al. 2020a). In order to investigate how changes in lincRNA-associated enhancer activity correlate with the expression of nearby PCGs, we compared the expression change of enhancer associated lincRNAs with that of the neighboring genes. There is a positive correlation between lincRNA associated enhancer expression dynamics and that of neighboring protein coding genes (**Figure 3.9C, D; Figure S6C, D**). The expression fold change of enhancer associated lincRNAs and its neighboring genes are always clustered together at the same time point (**Figure S6C**). Moreover, we plot the time points separately and observe moderate positive correlation (not high correlation values) between enhancer associated lincRNAs expression and the neighboring target genes (**Figure 3.9C, D**; **Figure S6D**). The trend holds true for all time points and it is more evident at the later time point (4_8) (**Figure 3.9D**). In order to confirm this trend, we also compare log2FC of the neighboring target genes of enhancer associated lincRNAs (la-g) and non-enhancer associated lincRNAs (na-g) (**Figure**

**3.9E**). We found log2FC of la-g are higher than of na-g and it's more evident in later flower development (4_8). Among target genes by AP1/SEP3 linked to enhancer associated lincRNAs, many flower related genes are included, such as *BRC1* (*BRANCHED 1*), *SUP* (*SUPERMAN*) and *PTL* (*PETAL LOSS*). Furthermore, there are 4 lincRNAs overlapping with 30 enhancers validated in Yan et al, 2019 (**Supplemental table S15**). For example, there is one lincRNA overlapping with one validated enhancer just upstream of the target gene *BRC1* (*BRANCHED 1*) which participates in axillary bud development (Xie et al. 2020). In summary, flower related lincRNAs are associated with enhancers and thereby contribute to floral gene regulation in *Arabidopsis*.

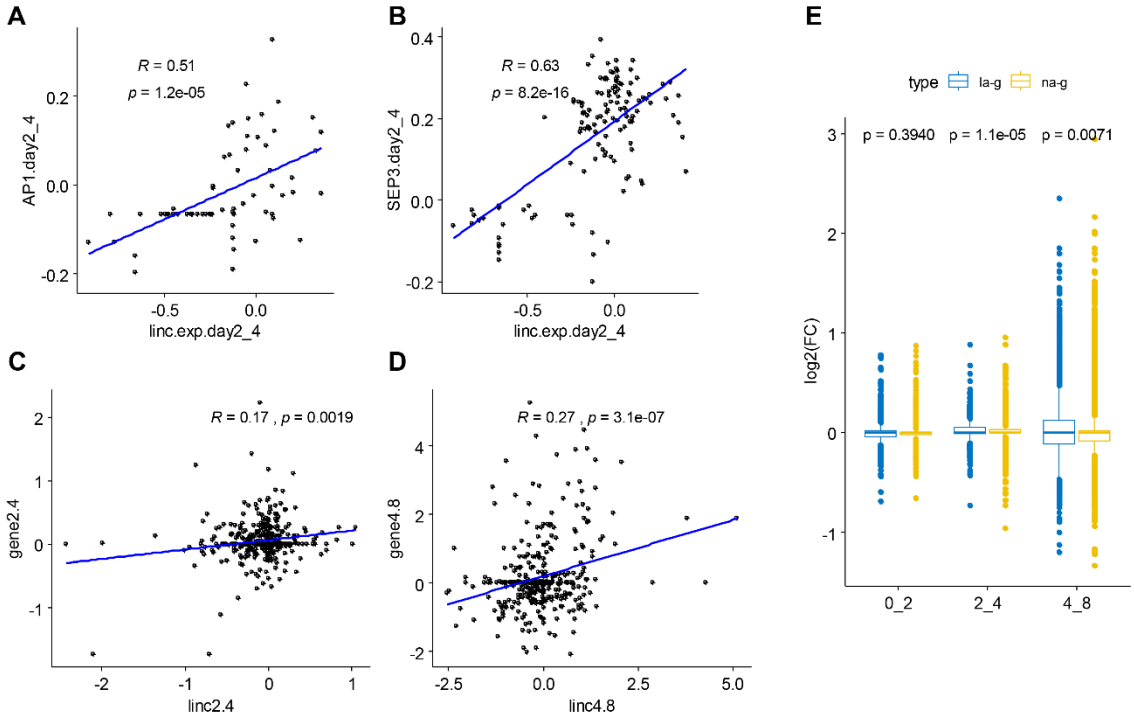



**Figure 3.9**: **Flower related lincRNAs are the components of floral gene regulatory network.** (A) Regression lines with pearson correlation coefficients between log2 (FC) change in expression of enhancers associated lincRNAs from day 2 to day 4 (linc.exp.day2_4, the same time point) and log2(FC) change in AP1 binding intensity from day 2 to day 4 (AP1.day2_4, the same time point). (B) Regression lines with Pearson correlation coefficients between log2(FC) change in expression of enhancers associated lincRNAs from day 2 to day 4 (linc.exp.day2_4, the same time point) and log2(FC) change in SEP3 binding intensity from day 2 to day 4 (SEP3.day2_4, the same time point). (C) Regression lines with Pearson correlation coefficients between log2 (FC) change in expression of enhancers associated lincRNAs from day 2 to day 4 (linc2.4, the previous time point) and log2 (FC) change in the neighboring target genes from day 2 to day 4 (gene2.4, the same time point). (D) Regression lines with Pearson correlation coefficients between log2 (FC) change in expression of enhancers associated lincRNAs from day 4 to day 8 (linc2.4, the previous time point) and log2 (FC) change in the neighboring target genes from day 4 to day 8 (gene4.8, the same time point). (E) Log2 (FC) of the neighboring target genes of

enhancer associated lincRNAs (la-g) and non-enhancer associated lincRNAs (na-g).

### 3.1.6 Functional investigation of enhancer associated lincRNA Aralnc.19175/linc-AP2 in floral gene regulatory network

Additionally, we tested Aralnc.19175 which was previously designated as *linc-AP2* (Gao et al. 2016) and was associated with the *AP2* gene. Aralnc19175 is overlapping with two enhancers (Chr4:17395366-17396138, Chr4:17396216-17396501) (**Figure 3.10A**) and bound by all seven master TF regulators including AP1, AP2, AP3, PI, AG, SEP3, and LFY (**Figure 3.10A, Figure 3.11E**). Intriguingly, the expression of the other neighboring gene AT4G36910 is not increasing too much (**Figure 3.11A**) while both the expression of lincRNA Aralnc19175 and the nearby *AP2* gene transiently induce 4 days after floral induction (**Figure 3.11B, C**). With the increasing expression of the lincRNA Aralnc.19175 (**Figure 3.11B**) and AP1/SEP3 binding intensity (**Figure 3.11F**), the chromatin accessibility of two lincRNAs associated enhancers is also increasing simultaneously (**Figure 3.11D**). Developmental TFs, such as AP1 and SEP3 bind to enhancers and induce expression of the enhancer-associated lincRNA Aralnc.19175 which then promotes openness of the chromatin in enhancers contributing to activation of the target protein coding gene *AP2* (**Figure 3.11G**). In summary, the data suggest that activity of the enhancer associated lincRNA Aralnc.19175/*linc-AP2* is linked to regulation of the neighboring gene *AP2* during flower development.

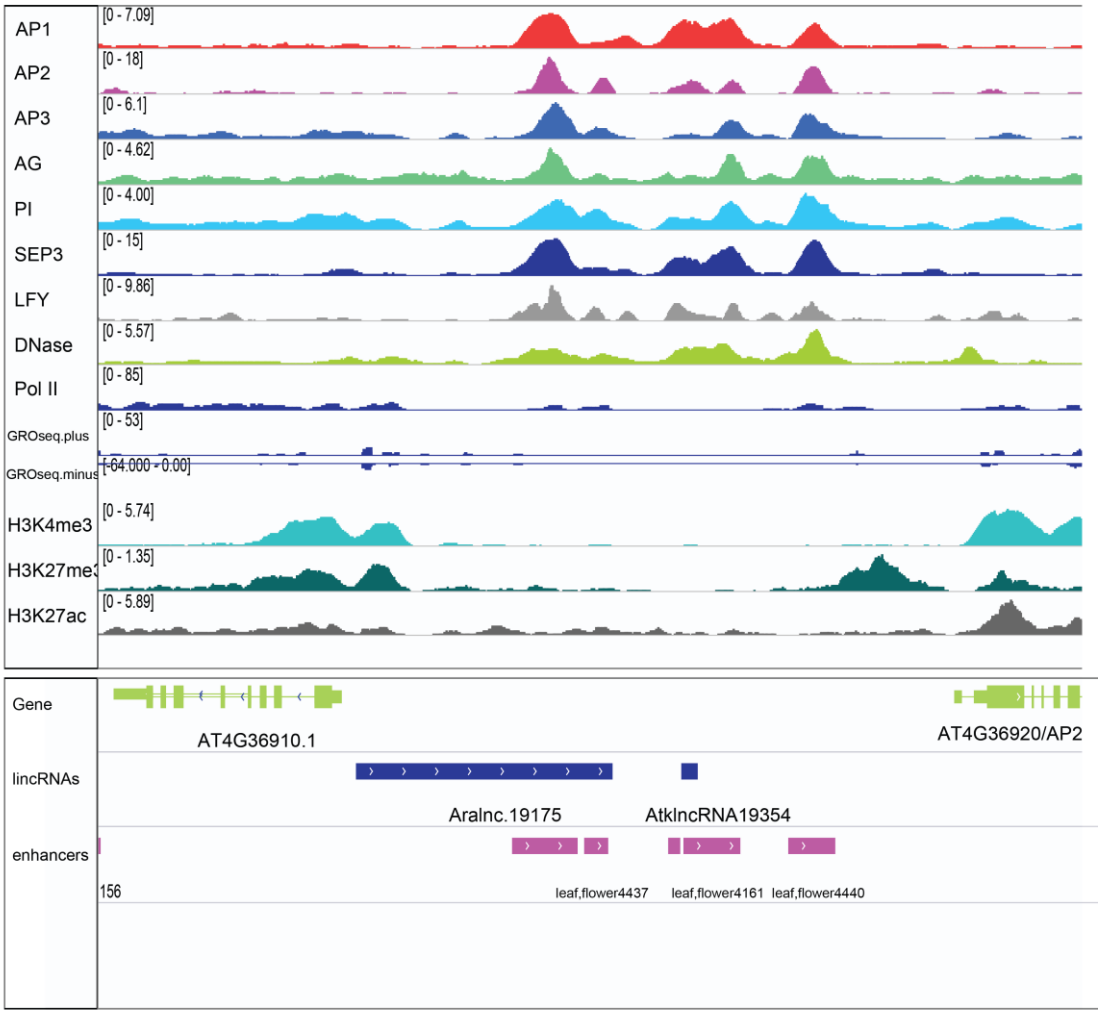



**Figure 3.10: TF binding and chromatin status at the Aralnc.19175/*linc-AP2* locus.** The example of lincRNA Aralnc.19175 locus targeted by master TFs and overlap with two enhancers (Chr4:17395366-17396138, Chr4:17396216-17396501).

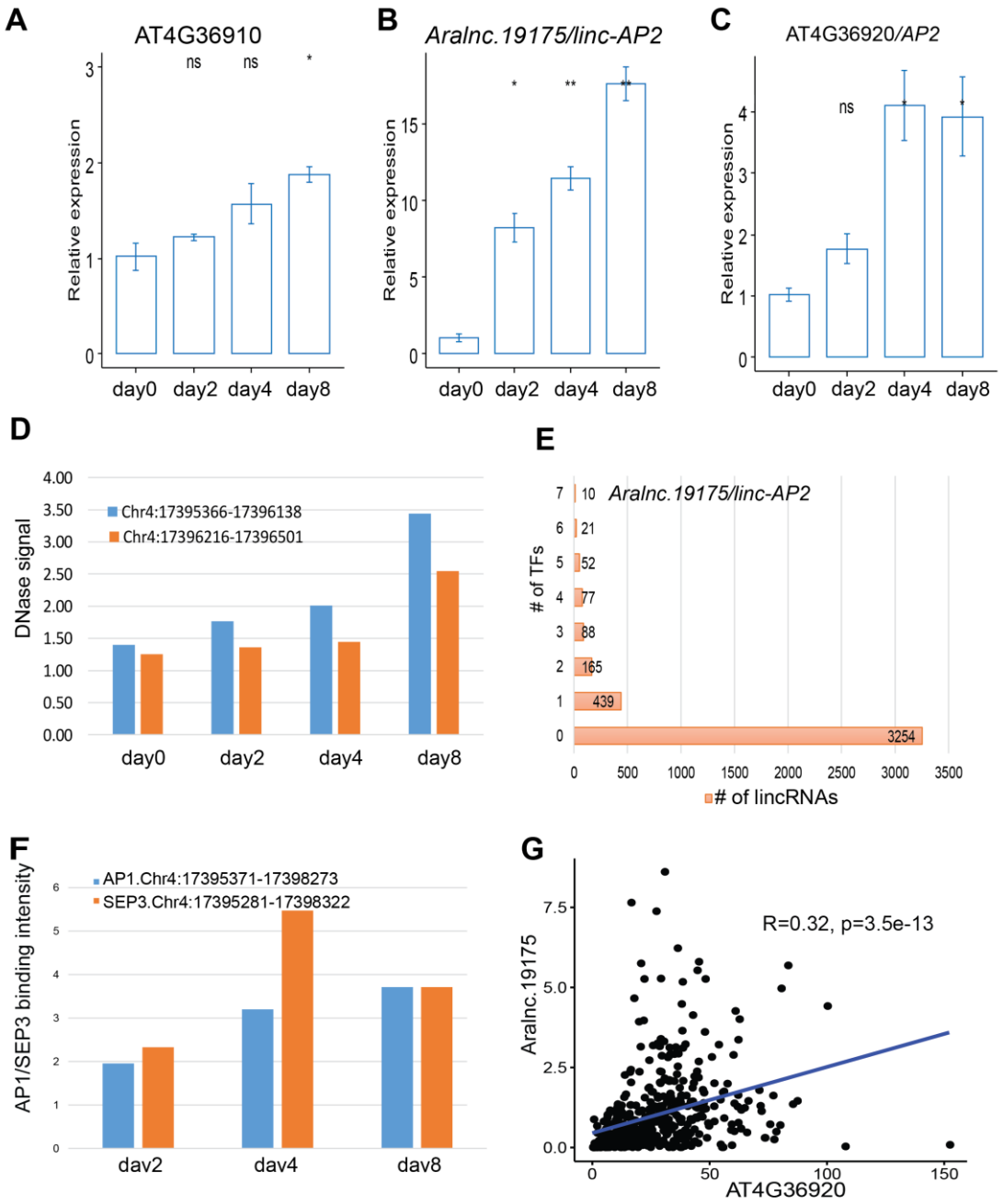



**Figure 3.11: Analysis of Aralnc.19175/*linc-AP2* activity.** (A) Expression pattern of the neighboring gene AT4G36910 across developmental stages. The expression of day0 was used as the reference for the calculation of p values. *, P-value <0.05; **, P-value <0.01; ***, P-value <0.001; ns, not significant. (B) Expression pattern of Aralnc.19175 across developmental stages. The expression of day 0 was used as the reference for the calculation of p values. *, P-value <0.05; **, P-value <0.01; ***, P-value <0.001; ns, not significant. (C) Expression pattern of the neighboring gene AT4G36920/*AP2* across developmental stages. The expression of day 0 was used as the reference for the calculation of p-values. *, P-value <0.05; **, P-value <0.01; ***, P-value <0.001; ns, not significant. (D) Chromatin accessibility patterns of two lincRNAs associated enhancers (Chr4:17395366-17396138, Chr4:17396216-17396501) across developmental stages. (E) The number of candidate lincRNAs regulated by # of TFs.

Aralnc.19175/*linc-AP2* are bound by all seven TFs. (F) The AP1/SEP3 binding intensity for binding sites at the lincRNA Aralnc.19175 loci. (G) The expression of Aralnc.19175/*linc-AP2* is correlated with the neighboring gene *AP2* demonstrated in tissues.

### 3.2 The evolutionary landscape of plant lincRNAs

#### 3.2.1 Genome wide identification of lincRNAs in 26 plant species reveals conserved characteristics of lincRNAs

To understand the evolution of plant lincRNAs and directly compare lincRNA transcripts from diverse plants, lincRNA transcripts were identified across 26 representative plant species including several non-flowering plants (**Figure 3.12A**). Harnessing a large number of RNA-seq datasets for developmental tissues and stages in each plant species (**Supplemental Table S16**), varying numbers of lincRNAs were identified (**Figure 3.12B, Supplemental Table S17**). We observed a higher number of lincRNAs for plant species with larger genomes (e.g. *Zea mays*). However, as demonstrated in other studies (Hezroni et al. 2015; Daish et al. 2014), direct comparison of lincRNA numbers in plant species was not easy because of the number and quality of the available RNA-seq data (**Figure S7A**), as well as the inherent heterogeneity in the sampled tissues. The viable size of the lincRNA repertoire size in different plant species may also be biologically meaningful (**Figure S7C**) although the proportion of expressed protein coding genes (PCGs) in each plant was relatively uniform (**Figure S7B**). Differences in sequencing depth, variable genome size, and assembly quality contributed to the overall differences of lincRNA numbers. We found that most lincRNAs had a single exon and one isoform (**Figure 3.13A, B**) irrespective of the species studied. Furthermore, features such as the maximum expression level and the size of lincRNAs were largely consistent across plant species, including non-flowering plants (**Figure 3.13C, D**), suggesting comparable quality of the identified lincRNAs and conserved features of lincRNAs in different plant species. For example, the expression levels of lincRNAs were consistently lower than that of PCGs (**Figure 3.13C**). For the plants (e.g. rice and soybean) with public available lincRNAs, we also compared the identified lincRNAs in this study with lincRNAs collected from publications and public databases (**Figure S7E**). It revealed the majority of lincRNAs identified in this study were novel ones, implying the incompleteness of lincRNAs in plants. Furthermore, we found that the number of lincRNAs in the 26 plant genomes was roughly in a linear relationship with their genome size (**Figure S7D**), and the genome sizes were well correlated with the number of lincRNAs identified per sample (**Figure 3.14**).

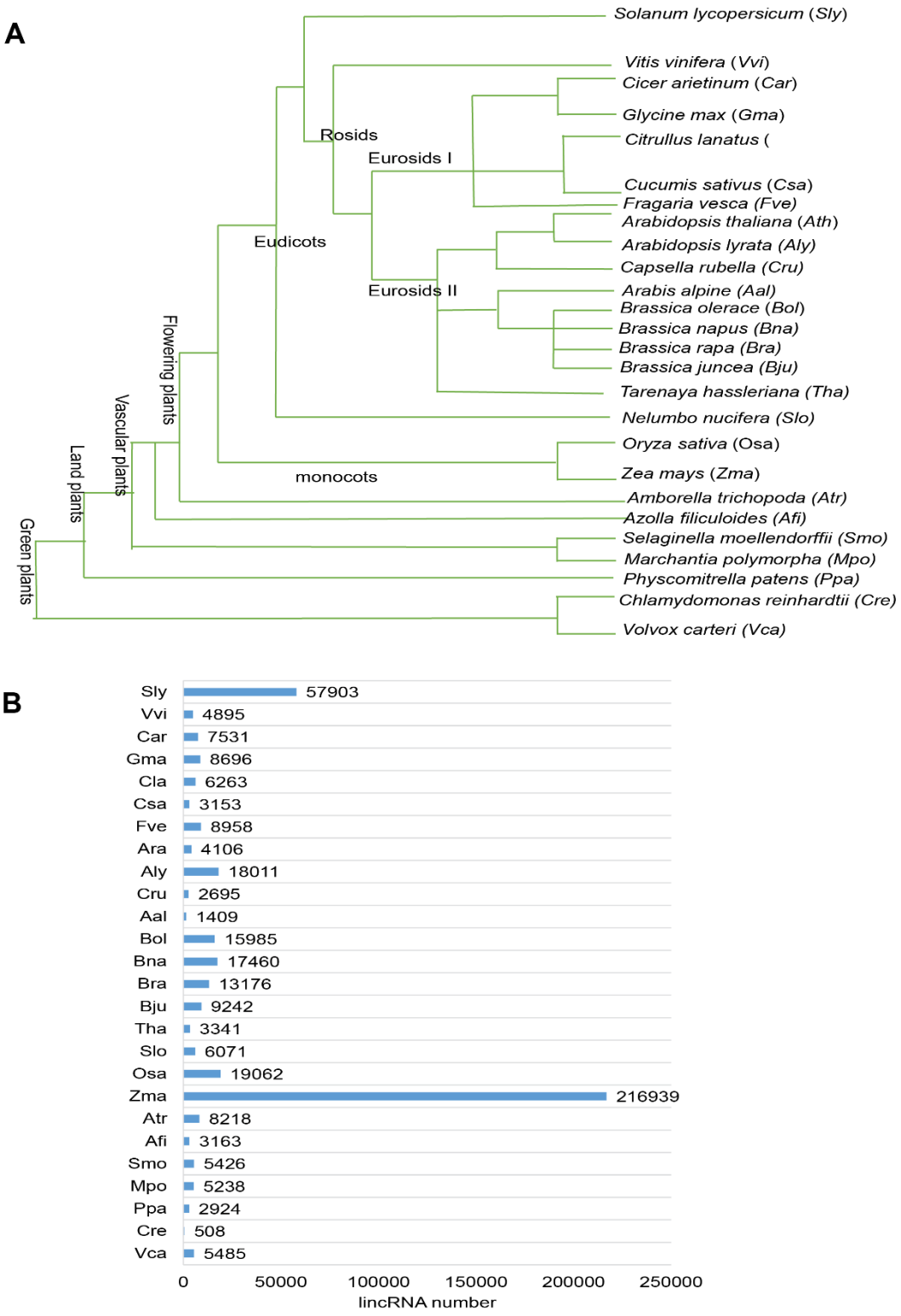



**Figure 3.12**: **Genome wide identification of lincRNAs across 26 plant species.** (A) The phylogenetic tree of 26 selected plant genomes including non-flowering plant species: *Cucumis sativus* (Csa), *Citrullus lanatus* (Cla), *Solanum lycopersicum* (Sly), *Vitis vinifera* (Vvi), *Cicer arietinum* (Car), *Glycine max* (Gma), *Fragaria vesca* (Fve), *Arabidopsis thaliana* (Ath), *Arabidopsis lyrata* (Aly), *Capsella rubella* (Cru), *Arabis alpine* (Aal), *Brassica oleracea* (Bol), *Brassica napus* (Bna), *Brassica rapa* (Bra), *Brassica juncea* (Bju), *Tarenaya hassleriana* (Tha), *Nelumbo nucifera* (Slo), *Oryza sativa* (Osa), *Zea mays*

(Zma), *Amborella trichopoda* (Atr), *Azolla filiculoides* (Afi), *Selaginella moellendorffii* (Smo), *Marchantia polymorpha* (Mpo), *Physcomitrella patens* (Ppa), *Chlamydomonas reinhardtii* (Cre) and *Volvox carteri* (Vca). (B) The number of lincRNAs identified in each plant species.

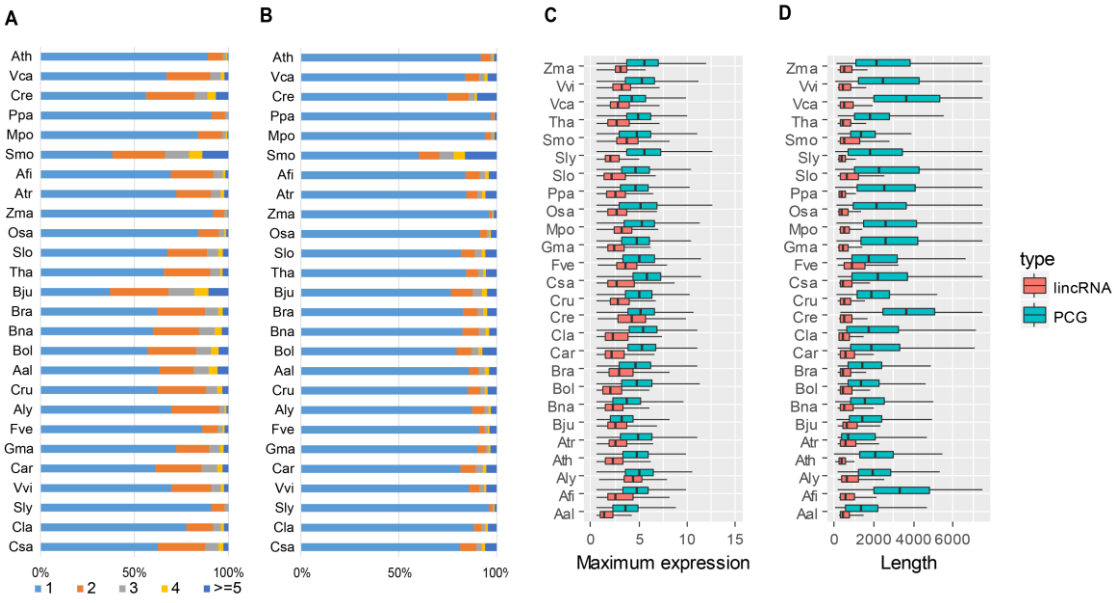



**Figure 3.13**: **Conserved features of lincRNAs across 26 plant species.** (A) The distribution of lincRNAs exon number in each plant species. (B) The distribution of lincRNAs isoform number in each plant species. (C) The maximum expression of both lincRNAs and protein-coding genes (PCGs) in each plant species. (D) The genomic length of both lincRNAs and protein-coding genes (PCGs) in each plant species.

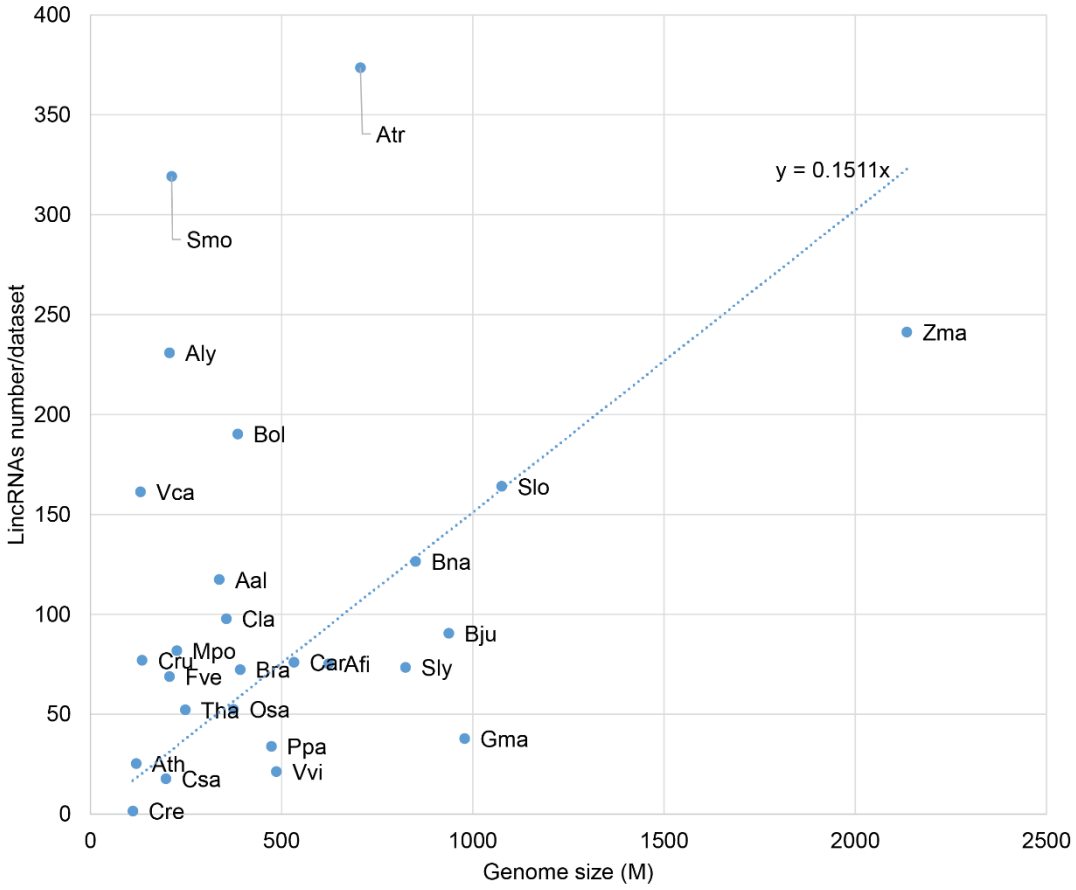



**Figure 3.14**: Genome size versus lincRNA number per dataset across 26 plant species.

### 3.2.2 Most lincRNAs are species-specific

Plant evolution has experienced several rounds of whole genome duplication (WGD), chromosome shuffle, and local duplication, all of them could contributed to multiply the copy numbers of lincRNAs (Qiao et al. 2019). Additionally, the ancestors of many flowering plants are polyploidy (Zhang et al. 2019a). Consequently, there are three main scenarios for lincRNAs: only a single copy in each of the selected plant species (defined as lincRNAs one2one family), a single copy in some of plant species and multiple copies in others (defined as lincRNAs one2many family), and multiple copies in all selected plant species (defined as lincRNAs many2many family) (**Figure 3.15A**). One typical scenario was illustrated in **Supplemental Figure S9**. In order to identify and classify lincRNA families, lincRNA sequences from the 26 plant species were compared pairwisely by the blast and their relationship was constructed using the graph clustering method MCL. In total, 18,937 lincRNAs families were identified based on similarity of lincRNAs sequences, including 10,355 (55%) of lincRNAs one2one family , 5,690 (30%) of lincRNAs one2many family and 2,892 (15%) of lincRNAs many2many family (**Figure 3.15A, Supplemental Table S18**). Most lincRNA families contained a single lincRNA from each plant species and were classified as lincRNAs one2one family type,

possibly suggesting rapid evolution of lincRNA loci. In the 6 non-flowering plants, *Azolla filiculoides* (Afi), *Selaginella moellendorffii* (Smo), *Marchantia polymorpha* (Mpo), *Physcomitrella patens* (Ppa), *Chlamydomonas reinhardtii* (Cre), and *Volvox carteri* (Vca), 2003 (11%) lincRNA families representing the most ancient ones were found. Homologous lincRNAs identified in different plant species usually shared only short patches (~60 nt) of sequence conservation (**Figure S8A**) with <10 mismatches within each patch (**Figure S8B**). Furthermore, a significant number of the lincRNAs identified here overlapped with the conserved non-coding sequences (CNS) reported in several previous studies (**Figure S8C, D, E**) (Van de Velde et al. 2016; de Velde et al. 2014; Haudry et al. 2013). In each plant species, the one2one family was the dominant family (**Figure 3.15B**). Most plant species except the *Brassicaceae* family had a low (<50%) percentage of homologous lincRNAs (**Figure 3.15C**) implying that most lincRNAs were species-specific due to rapid gain and/or loss during plant evolution, which is evident and demonstrated in the distribution of the number of homologous lincRNAs found in different species (**Figure 3.16A**). For example, in *Arabidopsis thaliana* (Ath), only 476 (476/4106, 11.6%) lincRNAs were highly conserved (defined as with homologous lincRNAs in at least six species) among the 313 families (257 many2many, 38 one2many, 18 one2one lincRNA family) (**Figure 3.16A**). Intriguingly, of the 476 highly conserved Ath lncRNAs, many of them were flanked by PCGs related to flowering and/or flower development, such as *FLO5, UFO1*, and *SACS3*, which presumably implies that these highly conserved lncRNA families may be also implicated in biological processes related to flower development. Despite lincRNAs displaying rapid sequence divergence compared to PCGs, 217 lincRNA families (94 one2one lincRNA family) identified in flowering plants (Angiosperms) had detectable sequence conservation and they made up a small subset of the lincRNAs that emerged over the past ~200 million years of flowering plant evolution (**Figure 3.16B, Figure S10A**). Together, identification and characterization of lincRNA families across the whole plant lineages revealed a rapid evolution of primary sequences of lincRNAs but still, a small portion of lincRNAs was preserved during the evolution of flowering plants.

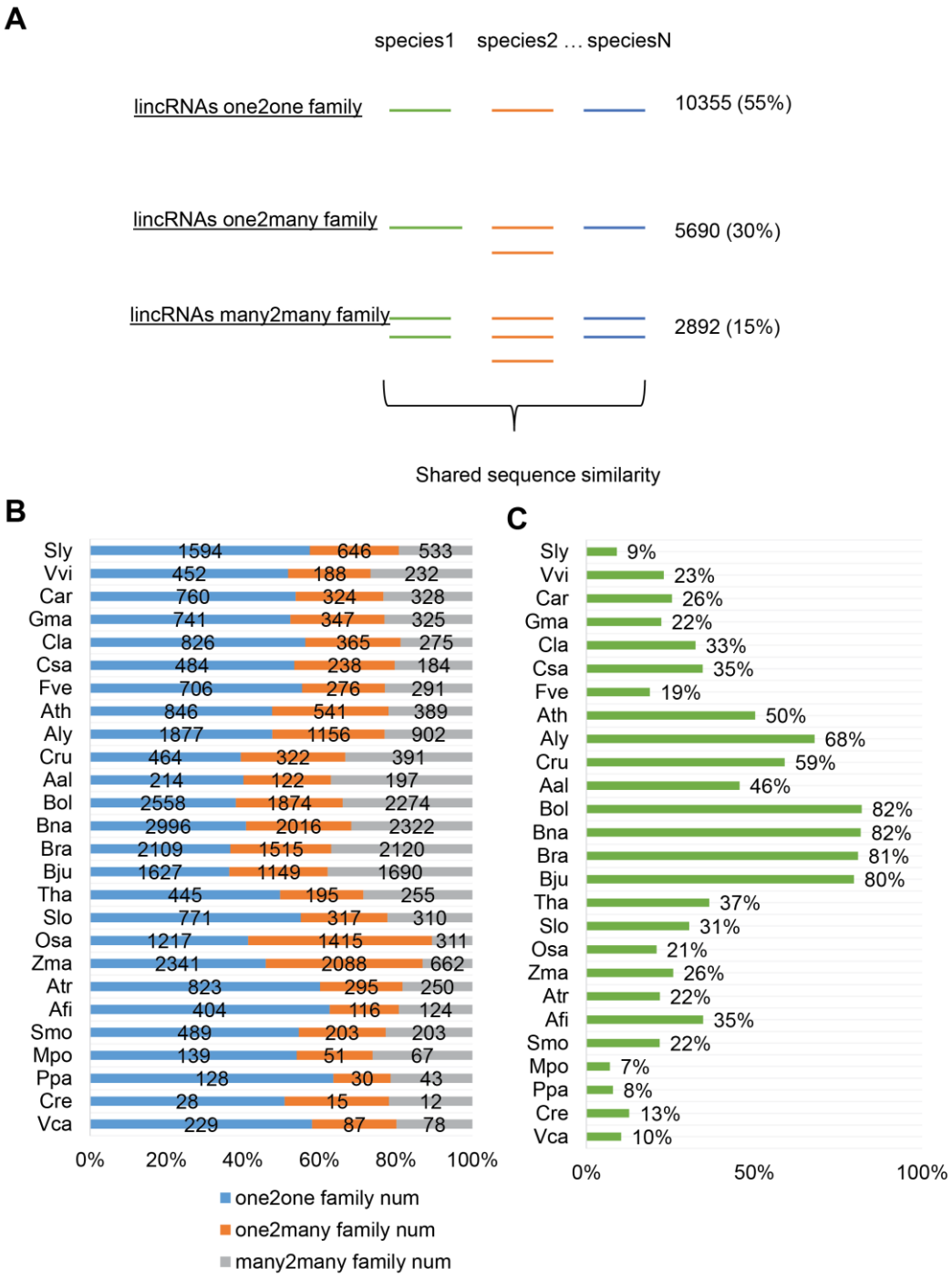



**Figure 3.15**: **Identification of lincRNA families by sequence similarity in plants including non-flowering plants.** (A) Three type lincRNAs families based on sequence similarity in plants: one2one family, one2many family, and many2many family. It shows the number and percentage of each type of lincRNAs family in plants. (B) The percentage of each type of lincRNAs family number in every plant species. (C) The percentage of the number of lincRNAs homolog in each plant species.

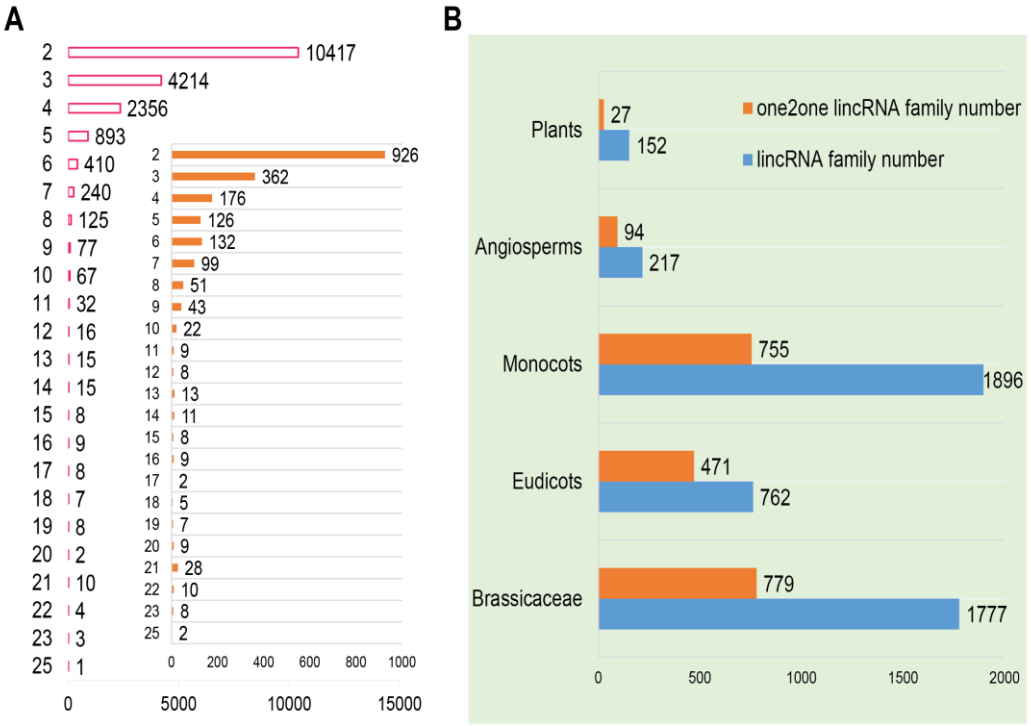



**Figure 3.16: Conservation of lincRNAs by sequence similarity in plants including non-flowering plants.** (A) The distribution of the number of species shared within one lincRNAs family. Insets: The distribution of the number of *Arabidopsis thaliana* lincRNAs shared with a certain number of species. (B) Identified conserved lncRNAs across different levels of lineages in plants (lincRNA evolutionary age groups: Plants, Angiosperms, Monocots, Eudicots, and *Brassicaceae*).

### 3.2.3 Transcriptional regulation of ancient lincRNAs in plants

Fast lincRNA evolution prohibits the identification of lincRNA homologs in distant species using the sequence homology based approach. It contributes to the smaller proportion of conserved lincRNAs in plants. In order to further understand the conservation of lincRNAs and the regulatory mechanism(s) underlying conserved lincRNAs, we compared the PhastCons scores of lincRNAs, which were calculated based on DNA sequence conservation metrics of 20 angiosperm plant genomes (Hupalo and Kern 2013) and the 1001 *Arabidopsis* genomes datasets (Alonso-Blanco et al. 2016), within different evolutionary age groups (EAGs) with that of PCGs. Four EAGs determined based on the number of lincRNAs homologous to those of *Arabidopsis thaliana* were used in comparison. They were Plants (n=71), Angiosperms (n=11), Eudicots (n=65), and *Brassicaceae* (n=556). The PhastCons scores decreased from the EAG Plants to the EAG *Brassicaceae* and the median conservation score of Plants (~0.34) was comparable with that of PCGs (~0.42) (**Figure 3.17A, C**). The SNP frequency (SNPs/100bp) was increased from the EAG Plants to the EAG Brassicaceae, but their SNP frequencies were all

higher than that of PCGs (**Figure 3.17B, D**), suggesting that potential purifying selection and relaxation of constraining may be the driving force responsible for the lower conservation of lincRNAs. Conservation of the upstream and downstream regions of lincRNAs were comparable with that of PCGs in both *Arabidopsis* and rice (**Figure S12**). Furthermore, the more conserved old lincRNAs (defined as those in the EAGs of Plants, Angiosperms, and Eudicots; **Figure 3.10A**) seemed to have higher expression levels (**Figure 3.18B**) and lower tissue specificity (**Figure 3.18C**) compared to that of young lincRNAs (defined as those in the EAG of Brassicaceae). Besides, the comparable expression levels between the old lincRNAs and PCGs implies conserved evolutionary selective pressure for these two groups of transcripts at the levels of transcriptional and chromatin regulation. We observed comparable frequencies of histone modifications and transcription factor (TF) binding sites in regions of lincRNAs, suggesting active regulation of lincRNAs (**Figure S11A, B**). Interestingly, several histone modifications (e.g. H3K9me2, H3K27me1 and H2A.W.6, and H3K27ac) and MADS TFs (the master regulators of flower development) were found to preferentially bind to the regulatory regions (1kb upstream/downstream) of lincRNAs in plants (**Figure 3.18D, E; Figure S13**) while a diffeent set of histone modifications (e.g. H3K36me2, H3K4me1, H3K4me2, H3K36me3, H3K18ac, H3K4me3, and H3K9ac) and TFs were found to be preferentially associated with PCGs (**Figure S11C, D**). For example~13% of old lincRNAs contained the SEP3 binding sites, five times higher than the same binding sites observed in PCGs (~2.5%) in *Arabidopsis thaliana*. The higher association of old lincRNAs with MADS TFs implied important functions of the old lincRNAs in the flower development. Supporting this notion, genome-wide study of ancient lncRNAs in tetrapods have found linkage between old lincRNA and homeobox TFs playing a role in embryonic development (Necsulea et al. 2014). High enrichment of H3K9me2, H3K27me1 and H2A.W.6 in old lincRNAs suggest association of conserved lincRNAs with heterochromatin regions enriched with transposable elements (TEs) (**Figure 3.18D**) while in contrast PCGs seemed to be more associated with active chromatin environment (**Figure S11D**). Taken together, these results demonstrated tight regulation of the highly conserved old/ancient lincRNAs by TFs or TEs.

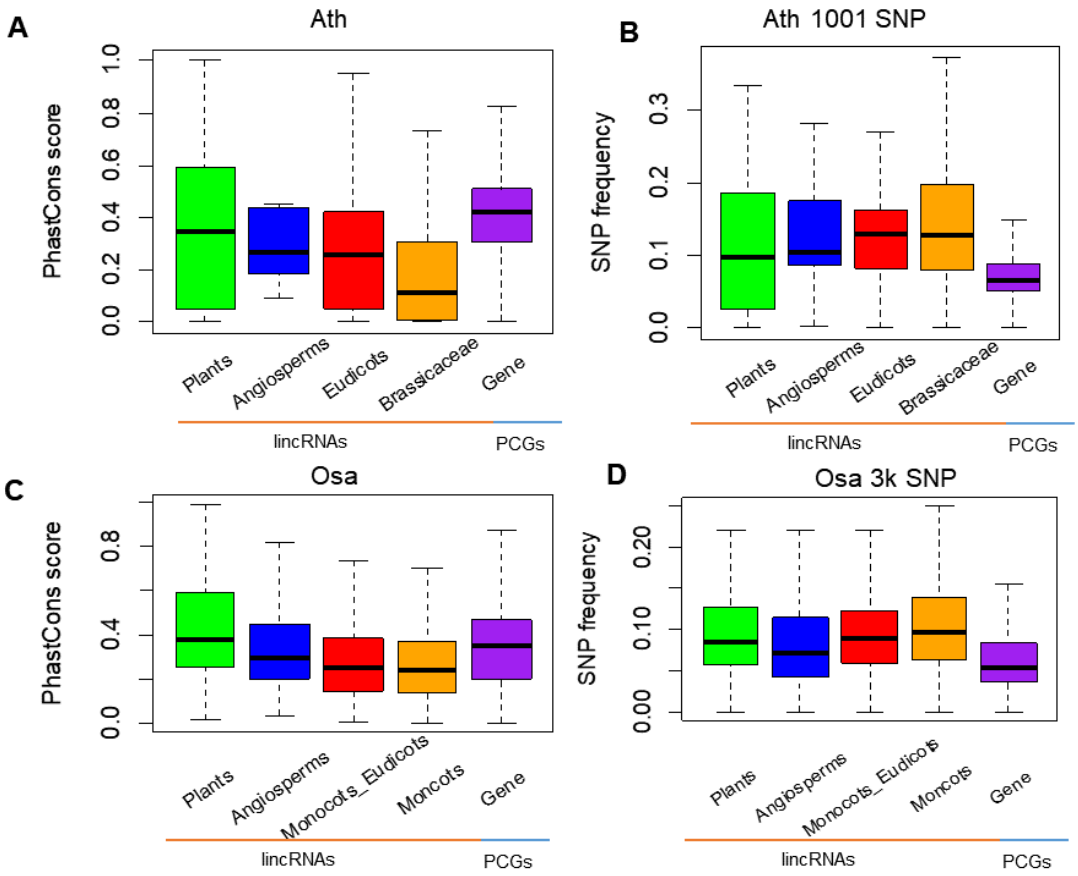



**Figure 3.17**. **Active regulation of ancient lincRNAs in plants.** (A) Sequence conservation (20 flowering plants genomes PhastCons scores) for lincRNA evolutionary age groups (Plants, n=71; Angiosperms, n=11; Eudicots, n=65; Brassicaceae, n=556), protein-coding genes (Gene, n=27655). (B) SNP frequency (SNPs/100bp) for lincRNA evolutionary age groups (Plants, n=71; Angiosperms, n=11; Eudicots, n=65; Brassicaceae, n=556), protein-coding genes (Gene, n=27655). (C) Sequence conservation for lincRNA evolutionary age groups (Plants, n=262; Angiosperms, n=111; Monocots_Eudicots, n=1023; Monocots, n=2482), protein-coding genes (Gene, n=27655). The PhastCons scores for Sequence conservation in rice are from the database PlantRegMap. (D) SNP frequency (SNPs/100bp) for lincRNA evolutionary age groups (Plants, n=262; Angiosperms, n=111; Monocots_Eudicots, n=1023; Monocots, n=2482), protein-coding genes (Gene, n=27655) in rice. SNPs in 3k rice are from SNP-Seek (https://snp-seek.irri.org/).

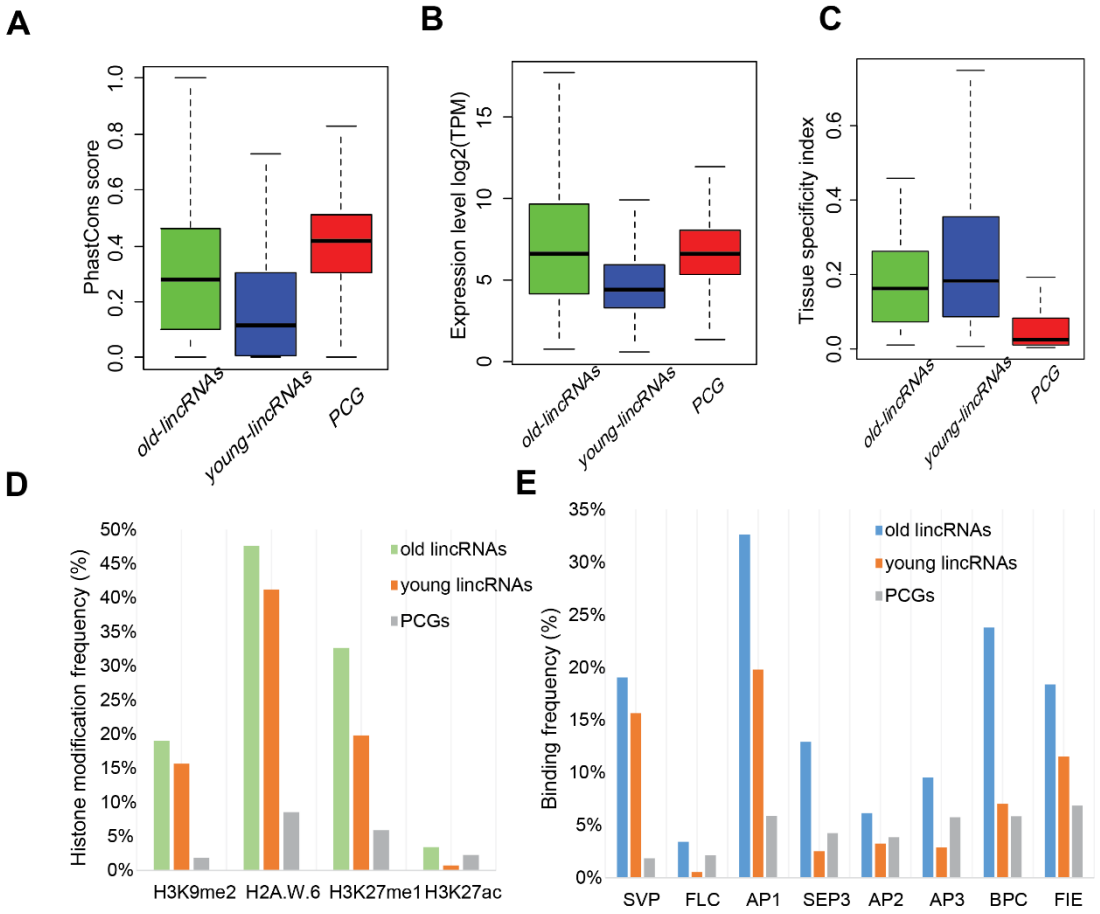



**Figure 3.18**. **Active regulation of ancient lincRNAs in plants.** (A) Sequence conservation (20 plants genomes PhastCons scores) for old and young lincRNAs (old lincRNAs (n=148): lincRNA evolutionary age groups in Plants, Angiosperms and Eudicots; young lincRNAs (n=566): lincRNA evolutionary age groups in *Brassicaceae*), protein-coding genes (Gene, n=27655). (B) The expression level for old and young lincRNAs. (C) Tissue specificity index for old and young lincRNAs. (D) Frequency of histone modification (H3K9me2, H2A.W.6, H3K27me1, and H3K27ac) in 1kb upstream/downstream regions of old, young lincRNAs and PCGs. (E) Frequency of binding sites for transcriptional factors (SVP, FLC, AP1, AP2, AP3, BPC, and SEP3) and FIE in 1kb upstream/downstream regions of old, young lincRNAs, and PCGs. Binding frequency (%) = percentage of old, young lincRNAs, and PCGs bound by TFs.

### 3.2.4 The expression pattern of lincRNAs suggests their high transcriptional turnover

In order to estimate the transcriptional activity of conserved lincRNAs across diverse plant species, we investigated transcription of *A. thaliana* lincRNAs and PCGs (used as a control) in other 25 plant species. We found that active transcription of lincRNAs homologous to those of *A. thaliana* was only evident in the plant species that are closely related to *A. thaliana* (**Figure 3.19A**). For example, about 40% of *A. thaliana* lincRNAs showed active transcription in *A. lyrata* while only ~1% of *A. thaliana* lincRNAs showed active transcription

in non-flower species. It is clear that even for *A. lyrate*, in which the highest transcription percentage of homologous lincRNAs was observed, the percentage was less than half of the transcription percentage of PCGs, whereas PCGs of *A. thaliana* were relatively constantly transcribed in other plant species (**Figure 3.19A**). These results suggest lincRNA expression pattern evolved quickly. To investigate the possible influence of tissues and samples on the results, we compared tissue specificity for the expressed lincRNAs in the three representative species of Brassicaceae: *Arabidopsis thaliana, Arabidopsis lyrata* (Aly)*, Capsella rubella* (Cru), for which data from equivalent tissues were available. When the hierarchical clustering method was used to cluster expression of lincRNAs from different samples of the three species, it is clear that different tissues from the same species were always clustered together (**Figure 3.19B**). However, only 12% of flower expressed lincRNAs in *A. thaliana* shared tissue specificity with the other two species (**Figure 3.19C**), similarly, ≤10% of Aly and Cru flower lincRNAs had their homologs expressed in Ath flowers, suggesthing a significant tissue-specificity of flower lincRNAs ($P<0.01$). The lincRNAs universally expressed in the flower tissues but not in other tissues in these three plant species would have a conserved function in flower development, such an example is shown in (**Figure 3.19D**).

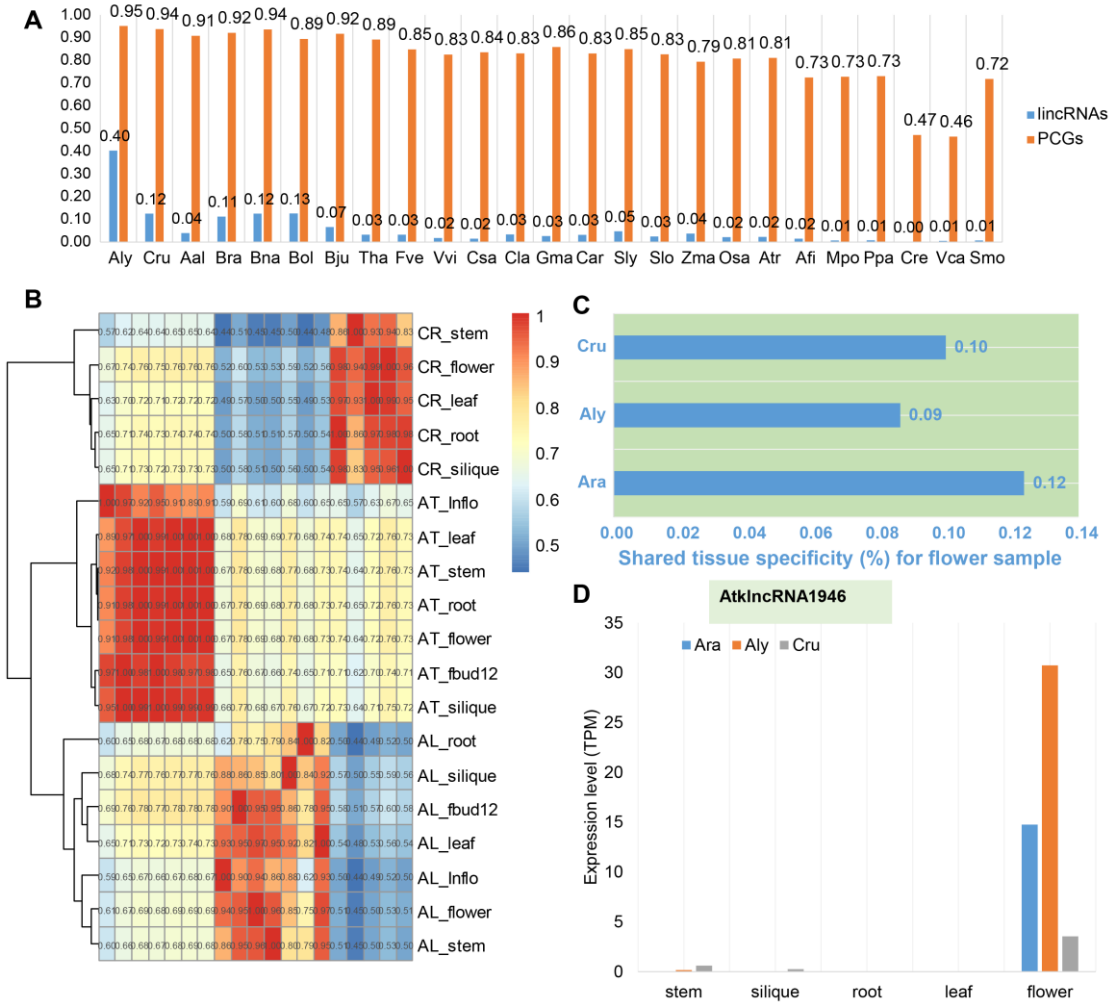



**Figure 3.19**: **The rapid transcriptional turnover during *Arabidopsis thaliana*, *Arabidopsis lyrata,* and *Capsella rubella* evolution.** (A) Percentage of *Arabidopsis thaliana* lincRNAs and protein-coding genes (PCGs) transcribed in other 25 plant genomes including non-flowering plants. (B) Hierarchical clustering of pairwise correlations for 277 lincRNA families during *Arabidopsis thaliana*, *Arabidopsis lyrata,* and *Capsella rubella* evolution. AT_: tissues in *Arabidopsis thaliana*; AL_: tissues in *Arabidopsis* lyrata; CR_: tissues in *Capsella rubella*. (C) The proportion of flower specific expressed lincRNAs in each species for which flower specificity are shared in *Arabidopsis thaliana*, *Arabidopsis lyrata,* and *Capsella rubella*. (D) A lincRNA (AtklncRNA1946) with conserved flower expression in *Arabidopsis thaliana*, *Arabidopsis lyrata,* and *Capsella rubella*.

To know whether similar tissue-specific expression of lincRNAs happened in monocots, we further investigated transcriptional profiles of lincRNAs in two representative monocot species, *Oryza sativa* and *Zea mays*, using RNA-seq datasets generated from different zones of elongating roots. Samples from the same species were grouped together (**Figure 3.20A**) and a relatively small percentage (<25%) of lincRNAs shared tissue specificity (**Figure 3.20B**). For example, only 18% of *Oryza sativa* lincRNAs shared tissue specificity with *Zea mays* in the

root meristematic zone. Nevertheless, some of the lincRNAs that were expressed in roots of both rice and maize, their expression patterns were quite constant in the two species, implying they may have a conserved function in root development. One such lincRNA is the pair Osalnc.47386 and Zmalnc.236427 that showed a decreasing expression pattern from root tip to differential zone (**Figure 3.20C**). A similar situation was evident for lincRNAs found in flower and reproductive tissues. LincRNAs were preferentially grouped together according to their origin, i.e. *Oryza sativa* or *Zea mays* (**Figure 3.21A)**. But like conserved lincRNAs in roots, lincRNAs expressed in specific tissues of both rice and maize were also found, such as the pair Zmalnc.293022 and Osalnc.21528 found in shoot apical meristem (**Figure 3.21B, C**).

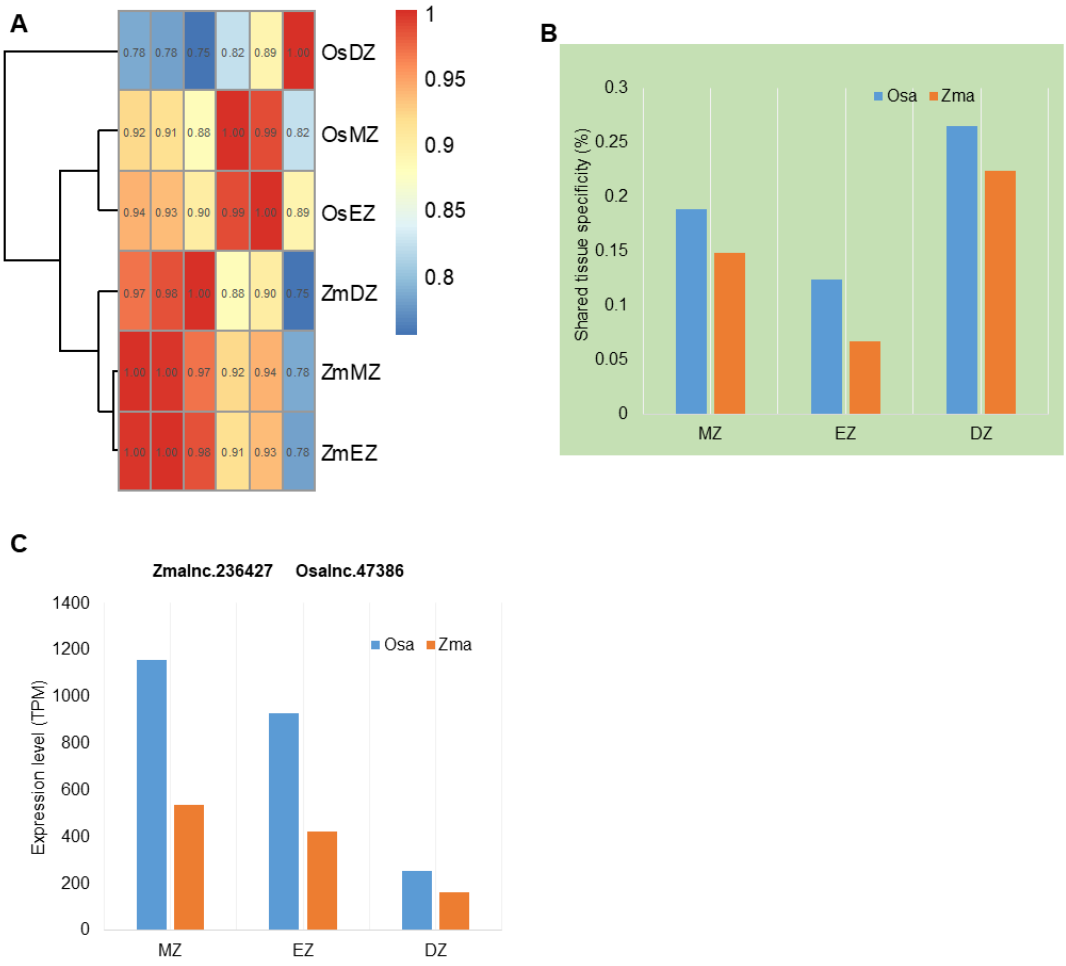



**Figure 3.20**: **The rapid change of tissue (root) specificity during *Oryza sativa* and *Zea mays* evolution.** (A) Hierarchical clustering of pairwise correlations for lincRNA families during *Oryza sativa* and *Zea mays* evolution. Os_: tissues in *Oryza sativa*; Zm_: tissues in *Zea mays*. (B) The proportion of tissue-specific expressed lincRNAs in each species for which the tissue specificity is shared in *Oryza sativa* and *Zea mays*. MZ: Root meristematic zone; EZ: Root elongation zone; DZ: Root differentiation zone. (C) LincRNAs (Zmalnc.236427/Osalnc.47386) with conserved expression in both *Oryza sativa* and *Zea mays*.

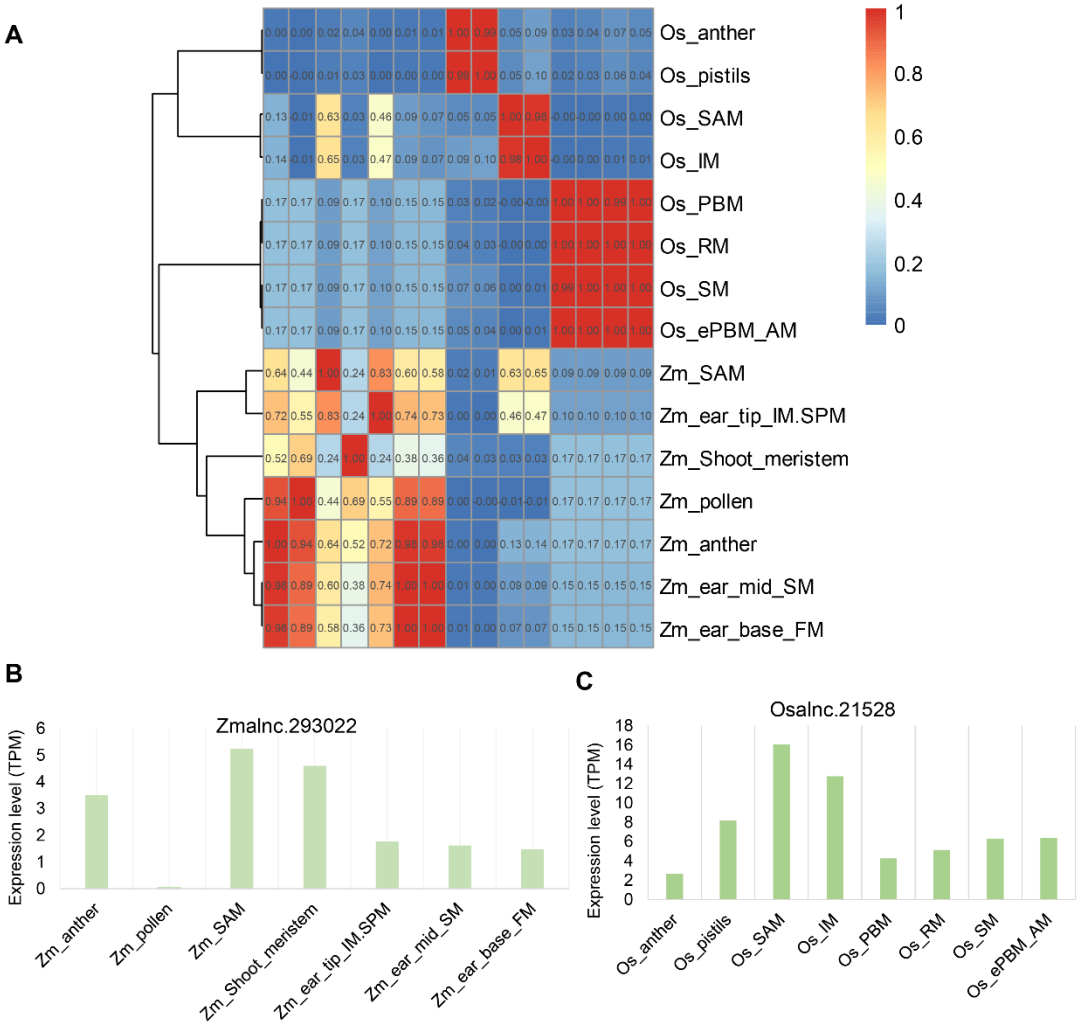



**Figure 3.21**: **The rapid change of tissue (meristem) specificity during *Oryza sativa* and *Zea mays* evolution.** (A) Hierarchical clustering of pairwise correlations for lincRNA families during *Oryza sativa* and *Zea mays* evolution. Os_: tissues in *Oryza sativa*; Zm_: tissues in *Zea mays*. (B) Zmalnc.293022 with expression in SAM datasets in *Zea mays*. (C) Osalnc.21528 with expression in SAM tissues in *Oryza sativa*.

### 3.2.5 Sequence-based homologous lincRNAs are largely not overlapping with synteny-based ones

Many putative lincRNA homologs cannot be detected through sequence similarity owing to their rapid sequence divergence; however, the genomic positions of such lincRNAs could be conserved during the evolution of plants. Therefore, syntenic relationship of PCGs could assist used to identify lincRNAs flanking the synteny PCGs despite little sequence similarity between the potential homologous lincRNAs (**Figure 3.22A**). Here, a syntenic block was defined when one or more of the three PCGs on each side of a given lincRNA and a total of 3

or more PCGs showed syntenic relatioship (Pegueroles et al. 2019). Using lincRNAs from *A. thaliana* as references, hundreds of syntenic lincRNAs were detected in other species of the *Brassicaceae* family and the number of synthenic lincRNAs dramatically reduced in plants that are evolutionary distant from *A. thaliana*. A small portion of lincRNAs showed both sequence and synteny homology (**Figure 3.22A**). For example, in *Arabidopsis lyrata,* 1592 lincRNAs were identified based on sequence homology*,* of which only 121 lincRNAs were detected by the synteny based method. Similar results were observed when using lincRNAs from other plants (e.g. *Arabidopsis lyrata*) as the reference (**Figure S14A**). Additionally, when using lincRNAs from other plants as the reference, the vast majority of lincRNAs identified based on the sequence similarity approach could not be identified based on syntenic relationship in the distant species (**Figure S14**). Between the two monocots, most lincRNAs only shared syntenic homology or sequence homology (**Figure S14C, D**). When considering the distribution of *A. thaliana* homologous lincRNAs detectable by both the sequence and synteny based approaches, we found that most of them were shared in the *Brassicaceae* family, especially in its nearest relative *Arabidopsis lyrata* (**Figure 3.22B**). Among the 199 homologous lncRNAs with both sequence based and synteny based homolog in *Arabidopsis thaliana*, most of them have one species supportive evidence (with both sequence based and synteny based homolog in the only one species) and multiple species supportive evidence lincRNAs (with both sequence based and synteny based homolog in the multiple species) (**Figure 3.22C**). LincRNAs supported by both sequence and syntheny were highly conserved in the *Brassicaceae* family. We found at least 34 such lincRNAs (**Figure S15**). One of them was illustrated in **Figure 3.23A**.

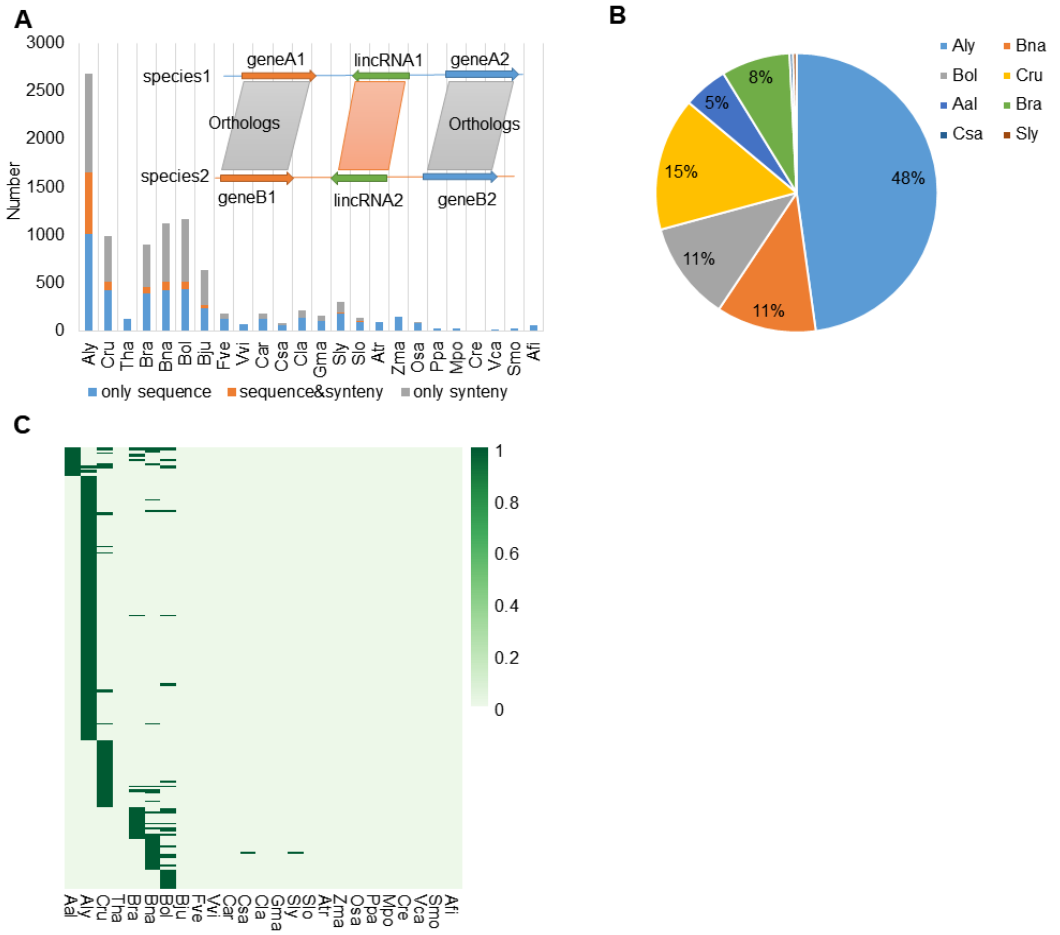



**Figure 3.22**: **Conservation of lincRNAs in the *Brassicaceae* family**. (A) The number of *Arabidopsis thaliana* sequence and/or synteny based homolog lincRNAs with other species. Insets: the approach for identification of synteny based homolog lincRNAs. (B) Distribution of homolog lincRNA pairs with other species which has the same sequence based and synteny based homolog in *Arabidopsis thaliana*. Most homolog lincRNAs in *Arabidopsis thaliana* are shared in the *Brassicaceae* family, especially in the nearest species *Arabidopsis lyrata*. (C) Among 199 homolog lncRNAs in *Arabidopsis thaliana*, most of them have one species supportive evidence (with both sequence based and synteny based homolog). Multiple species supportive evidence lincRNAs, one lncRNA in *Arabidopsis thaliana* has homolog (both sequence based and synteny based homolog) lincRNAs in multiple species.

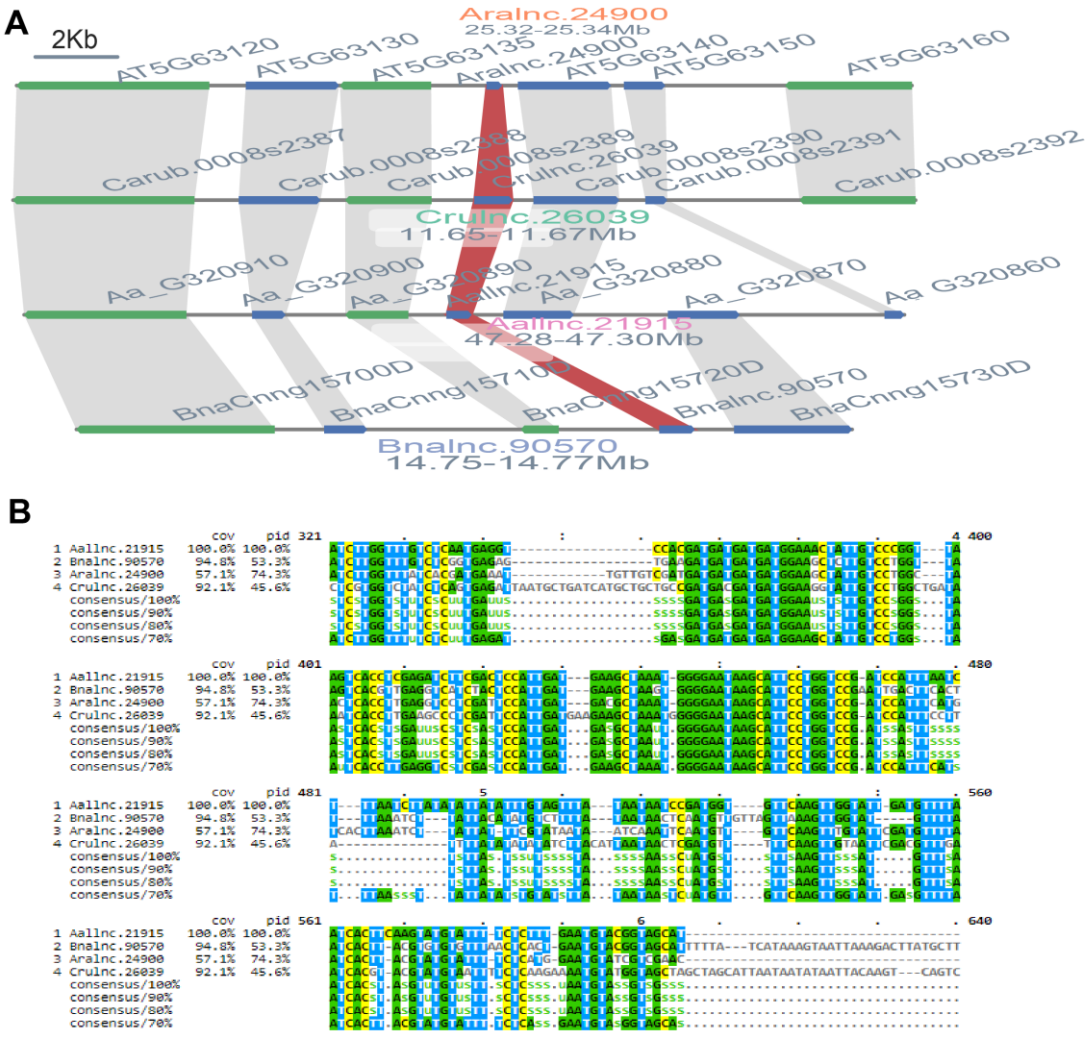



**Figure 3.23**: **Conservation of Aralnc.24900 in the *Brassicaceae* family.** (A) Aralnc.24900 in *Arabidopsis thaliana*. (B) Alignment of homolog lincRNAs in *Arabidopsis thaliana* (Aralnc.24900), *Arabis alpine* (Aallnc.21915), *Capsella rubella* (Crulnc.26039), and *Brassica napus* (Bnalnc.90570).

### 3.2.6 Synteny and gene network based functional characterization of conserved lincRNAs

LincRNAs regulate gene expression by in *cis* or in *trans* mechanism (Marchese et al. 2017). To investigate potential function of lincRNAs, we used neighboring PCGs to infer in *cis* function of lincRNAs while used co-expression to infer their potential in *trans* functions. Some of the lincRNAs were flanked by conserved genes related to flowering pathways, such as *BRC1/TB1* (**Figure S15, Figure S16A**), *AG* (**Figure S16B**), *LFY* (**Figure S16C**), *SEP1* (**Figure S16D**), *FT/ZCN* (**Figure S16E**) and *SOC1* (**Figure S16F**) suggesting that these lincRNAs may potentially function in *cis* to regulate the conserved functions of their neighboring PCGs.

We also used the expression levels of PCGs and lincRNAs of multiple samples to compute their co-expression relationship using WGCNA in the following 7 representative plants species: *A. thaliana*, *A. lyrata*, *C. rubella*, *B. napus*, *O. sativa*, *Z. mays,* and *M. polymorpha*. In each plant

species, several co-expression modules were identified and the PCGs included in each module were subjected to GO enrichment analysis. Here, we presented the results by using flowering related modules as examples. The PCGs within the module Ara.Module36 were enriched with GO terms related to flower development (qvalue= 1.91e-20), meristem development (qvalue= 1.60e-18), meristem maintenance (qvalue= 7.54e-14), and these PCGs had strong expression levels in meristems and flowers (**Figure S17A**). We hypothesized that the lincRNAs within each module would function in the same pathway(s) as their co-expressed PCGs. For instance, in the module of Ara.Module36, a lincRNA, Aralnc.24900, well conserved in Aal (Aallnc.21915), Bna (Bnalnc.90570) and Cru (Crulnc.26039) (**Figure 3.23A**), was co-expressed with several flowering related genes, including *LFY*, *STM*, *FUL,* and *AP1* (**Figure S17B, Figure 3.24A, B**). Aligning these lincRNA sequences found many conserved motifs that may potentially serve as sites for RNA binding or other functionality (**Figure 3.23B**). Indeed, some of those were binding sites of several master regulatory TFs (e.g. *LFY*, *AP1*, and *SEP3*) of flower development (**Figure 3.25B, C**). Additionally, the homologous lincRNAs of Aralnc.24900 were also found in flower related modules in Cru (Cru.Module31) (**Figure 3.24A, C**) and Bna (Bna.Module116) (**Figure 3.24A, D**), in which they were co-expressed with the same sets of flower related genes identified in Ara.Module36 (**Figure 3.25A**).

**A**

| name | module | Alldegrees.kTot | Alldegrees. Within | MM.kM | kME.pvalu |
|---|---|---|---|---|---|
| Aralnc.2490 | Ara.Module36 | 1.28 | 0.97 | 0.58 | 3.23E-93 |
| Bnalnc.9057 | Bna.Module11 | 179.63 | 159.56 | 0.80 | 7.89E-30 |
| Crulnc.2603 | Cru.Module31 | 425.07 | 401.63 | 0.87 | 1.63E-11 |
| LFY | Ara.Module36 | 11.08 | 10.60 | 0.82 | 7.14 E-257 |
| STM | Ara.Module36 | 9.96 | 8.93 | 0.82 | 1.20E-250 |
| FUL | Ara.Module36 | 5.91 | 2.77 | 0.61 | 1.41E-105 |
| AP1 | Ara.Module36 | 5.17 | 2.09 | 0.58 | 1.82E-92 |

**B**

**Ara.Module36**

| GOid | GO_desc | FDR |
|---|---|---|
| GO:0048367 | shoot system develc | 1.35E-23 |
| GO:0090567 | reproductive shoot s development | 1.14E-18 |
| GO:0007275 | multicellular organism development | 5.13E-18 |
| GO:0009908 | flower development | 5.23E-18 |
| GO:0032502 | developmental process | 2.72E-16 |
| GO:0048507 | meristem development | 4.17E-16 |
| GO:0048856 | anatomical structure development | 7.98E-16 |
| GO:0048731 | system development | 8.51E-16 |
| GO:0009791 | pos embryonic develc | 2.05 E-13 |
| GO:0009888 | tissue development | 4.39E-13 |
| GO:0099402 | plant organ development | 2.19E-12 |
| GO:0010073 | meristem maintenance | 1.45E-11 |
| GO:0061458 | reproductive system development | 1.63E-11 |
| GO:0048608 | reproductive structure development | 1.89E-11 |

**C**

**Cru.Module31**

| GOid | GO_desc | FDR |
|---|---|---|
| GO:009040( | pollen tube | 4.96E-05 |
| GO:008009: | regulation of pollen tube growth | 4.96E-05 |
| GO:000986( | pollen tube growth | 0.000437203 |
| GO:004823: | pollen sperm cell differentiation | 0.000575364 |
| GO:001058: | floral meristem determinacy | 0.002275406 |
| GO:001058- | pollen exine formation | 0.00739362 |
| GO:009040- | pollen tube tip | 0.017622811 |

**D**

**Bna.Module116**

| GOid | GO_desc | FDR |
|---|---|---|
| GO:0010254 | nectary development | 0.000798156 |
| GO:0010582 | floral meristem determinacy | 0.016015571 |
| GO:0009934 | regulation of meristem structural organization | 0.039015137 |
| GO:0048439 | flower morphogenesis | 0.041749133 |
| GO:0009911 | positive regulation of flower development | 0.04775327 |

**Figure 3.24: The functionality of Aralnc.24900 by the co-expression network.** (A) Network

characteristics of homolog lincRNAs within the module Ara.Module36 in *Arabidopsis thaliana*, Cru.Module31 in *Capsella rubella* and Bna.Module116 in *Brassica napus*. (B) GO annotation of the module Ara.Module36 in *Arabidopsis thaliana.* (C) GO annotation of the module Cru.Module31 in *Capsella rubella*. (D) GO annotation of the module Bna.Module116 in *Brassica napus.*

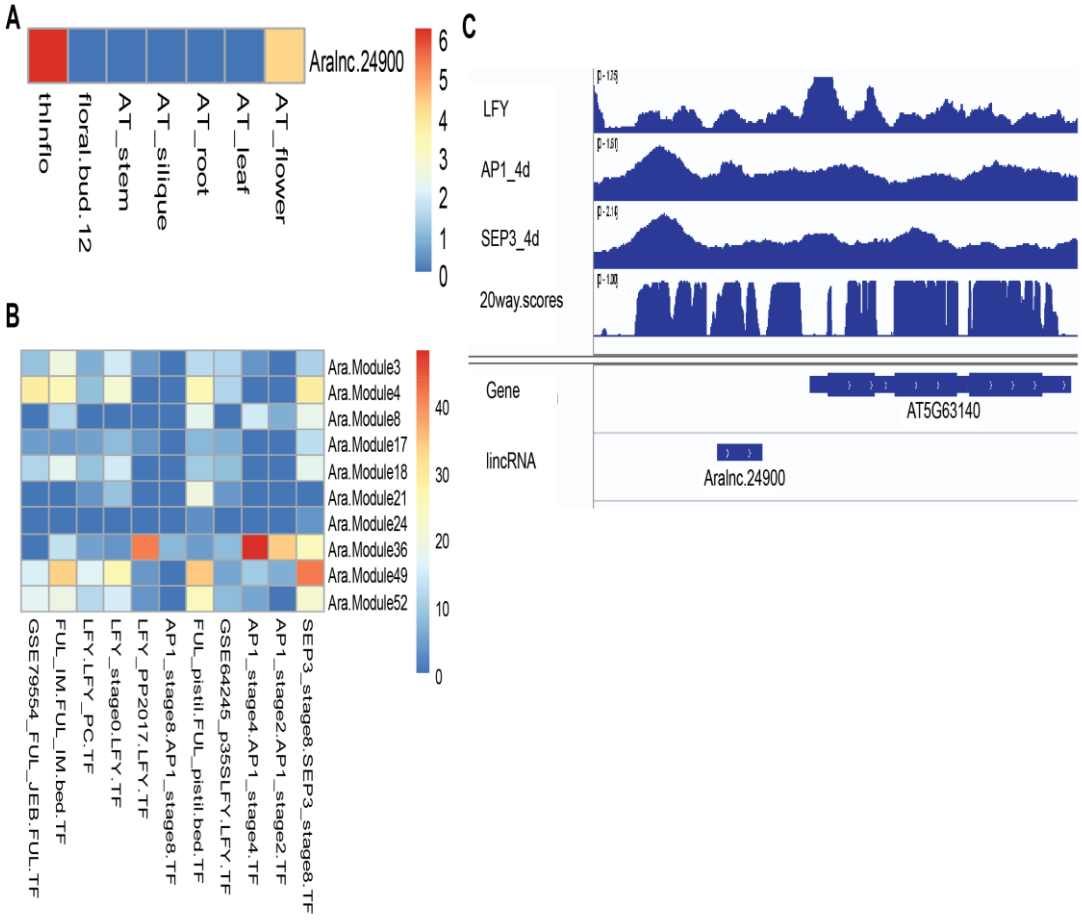



**Figure 3.25: The functionality of Aralnc.24900 by the co-expression network.** (A) The expression pattern of Aralnc.24900 in *Arabidopsis thaliana.* (B) The enrichment of PCGs in flower-related modules for target genes of AP1, SEP3, LFY, and FUL determined by Chip-seq. (C) The coverage map for *LFY*, *AP1,* and *SEP3* around Aralnc.24900 in *Arabidopsis thaliana.* 20way.scores: the track for PhastCons scores.

We also investigated the potential functions of the lincRNAs conserved in *Oryza sativa* and *Zea mays* (**Figure 3.26**). Based on sequence similarity, 235 and 2879 lincRNAs were identified in *Oryza sativa* and *Zea mays*, respectively. Of these conserved lincRNAs, only 7 were also identified based on the synteny homology approach. In order to infer the functionality of these conserved lincRNAs, for each species co-expression networks involving lincRNAs and PCGs were constructed by WGCNA, from which several co-expression modules

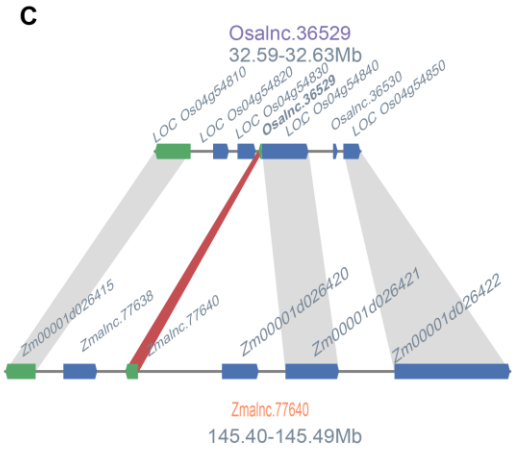
were identified. Similar to the results achieved in dicots, we also found co-expressed module enriched with PCGs related to flower/meristem development in both rice and maize (**Figure S19, S20)**, implying similar functions of the lincRNAs and PCGs of the corresponding modules identified in the two plant species (**Figure S19, S20**). For instance, Osalnc.36529 of *Oryza sativa* and Zmalnc.77640 of *Zea mays* were found in the syntenic region (**Figure 3.26C**) in the flower related module Osa.module81 (**Figure 3.26A**) and Zma.module3 (**Figure 3.26B**), respectively. Osa.module81 was enriched with GO:0016049 (cell growth, 1.39E-13), GO:0009856 (pollination, 6.35E-11), and GO:0030154 (cell differentiation, 0.0000236) (**Figure 3.26A**), and Osalnc.36529 was highly expressed in anthers and pistils (**Figure 3.26D**). Correspondingly, Zma.module3 had functions related to GO:0019953 (sexual reproduction, 1.98E-13), GO:0044703 (multi-organism reproductive process, 2.05E-13), and GO:0071555 (cell wall organization, 2.84E-13) (**Figure 3.26B**) and Zmalnc.77640 was highly expressed in pollens (**Figure 3.26E**). Based on these results, we conclude that highly conserved lincRNAs usually have similar functionality in different plants and act coordinately with their co-expressed PCG partners.

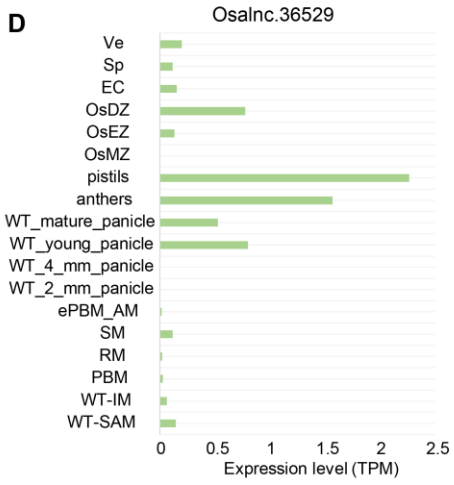
**A** Osa.module81

| GOid | GO_desc | FDR |
|---|---|---|
| GO:0016049 | BP cell growth | 1.39066E-13 |
| GO:0009856 | BP pollination | 6.35212E-11 |
| GO:0016043 | BP cellular component organization | 2.51389E-08 |
| GO:0009653 | BP anatomical structure morphogenesis | 3.09541E-08 |
| GO:0008152 | BP metabolic process | 2.34814E-06 |
| GO:0009987 | BP cellular process | 2.89501E-06 |
| GO:0030154 | BP cell differentiation | 2.36377E-05 |

**B** Zma.module3

| GOid | GO_desc | FDR |
|---|---|---|
| GO:0019953 | BP sexual reproduction | 1.97582E-13 |
| GO:0044703 | BP multi-organism reproductive process | 2.05051E-13 |
| GO:0071555 | BP cell wall organization | 2.83631E-13 |
| GO:0045229 | BP external encapsulating structure organization | 1.00262E-12 |
| GO:0055085 | BP transmembrane transport | 2.61885E-11 |
| GO:0098655 | BP cation transmembrane transport | 6.12768E-11 |
| GO:0034220 | BP ion transmembrane transport | 6.12768E-11 |



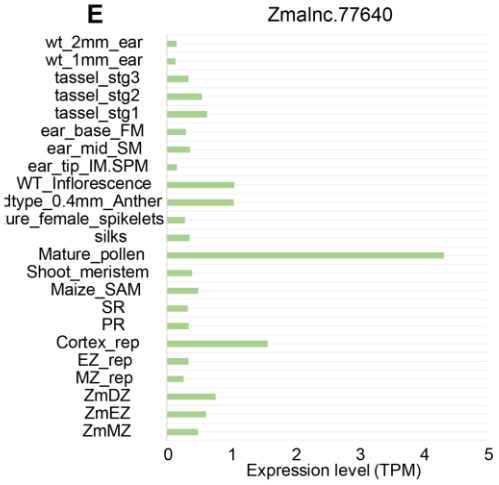



**Figure 3.26**: Conservation of lincRNAs in the *grass* family. (A) GO annotation of the flower-related module Osa.module81 in *Oryza sativa.* (B) GO annotation of the flower-related module Zma.module3 in *Zea mays.* (C) Osalnc.36529 in *Oryza sativa* is the syntenic region with Zmalnc.77640 in *Zea mays.* (D) The expression pattern of Osalnc.36529 in *Oryza sativa.* (E) The expression pattern of Zmalnc.77640 in *Zea mays.*

### 3.2.7 TEs drive evolutionary stabilization of lincRNAs in plants

A large portion of lincRNAs have been found to be overlapping with transposable elements (TEs) in human and mouse (Kapusta et al. 2013). The number of TEs varies significantly in different plant species and many plant genomes such as *Zea mays* possess high contents of TEs. TEs have been demonstrated to have significant impacts on plant genome evolution. But how about their role in the evolution of plant lincRNAs? To address this question, we checked overlapping between the identified lincRNAs and TEs in each plant genome using the bedtools. We found that TEs were highly associated with lincRNAs and most lincRNAs seemed to be derived from TEs (**Figure 3.27A**). The highest association between lincRNAs and TEs was found in maize, but interestingly *A. lyrata* (Aly) had a higher rate of association (86.2%) than most of other plants despite its modest TE content (~31.1%) in the genome. LincRNAs associated with TEs have been suggested to be rewired by the associated TEs (Lv et al. 2019). We found that lincRNAs of different plant families were linked with different types of TEs (**Figure 3.28A**). For example, in the *Brassicaceae* family, DNA/Helitron seemed to be the dominant ones, however, in monocots, the dominant TEs were LTR/Gypsy. In order to understand whether and how TEs drove lincRNA evolution in plants. We compared the percentage of the conserved lincRNAs associated with TEs in the representative eudicot (Ath) and monocot (Osa) species, and in different evolutionary age groups of the species. In *Arabidopsis thaliana*, species-specific (Ath-specific) lincRNAs were depleted of TEs (**Figure 3.27B**) compared with those conserved in different evolutionary age groups, while in *Oryza sativa,* the proportion of species-specific lincRNAs associated with TEs was similar to that observed in other evolutionary age groups (**Figure 3.27C**). The lincRNAs conserved in rice were more likely to be associated with TEs than those conserved in *Arabidopsis*, a phenomenon that might be related to the transposition mechanisms of the TEs involved (retrotransposons vs transposons) (**Figure 3.16B, C**). We further used lincRNAs conserved in Ath, Aly, and Cru and the remaining lincRNA in each of three species to compare their association with TEs. No matter which species, Ath, Aly, or Cru, conserved lincRNAs always had a higher fraction associated with TEs compared to the non-conserved ones (**Figure 3.28B**). An example of a lincRNA conserved in Ath, Aly, and Cru and their associated TEs was illustrated (**Figure 3.28D**). Similar result was also observed in the monocot species Osa and Zma although with a less

difference (**Figure 3.28C**). In summary, compared to PCGs, lincRNAs are likely to be associated with TEs, which may be the driving force for the evolution of lincRNAs or the origin of lincRNAs.

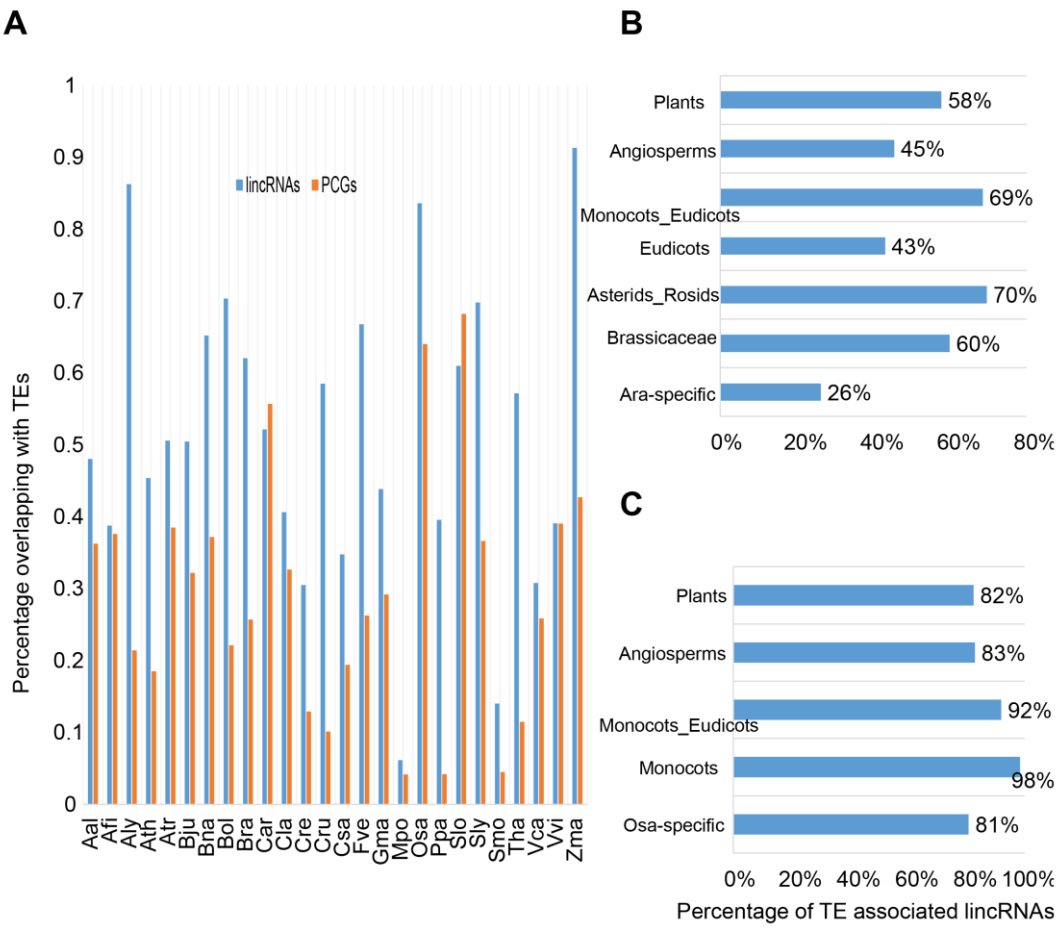



**Figure 3.27**: **Transposable elements (TEs) are associated with lincRNAs.** (A) The fraction of lincRNAs and genomes overlapping with TEs. (B) Different evolutionary age groups (Plants (n=71), Angiosperms (n=11), Monocots_Eudicots (n=242), Eudicots (n=65), Asterids_Rosids (n=135), *Brassicaceae* (n=556), Ara-specific (n=2044) in decreasing order of evolutionary age) of lincRNAs overlapping with TEs in *Arabidopsis thaliana*. (C) Different evolutionary age groups (Plants (n=262), Angiosperms (n=111), Monocots_Eudicots (n=1023), Monocots (n=2482), Osa-specific (15073) in decreasing order of evolutionary age) of lincRNAs overlapping with TEs in *Oryza sativa*.

A

| Species | Top3 lincRNA associated TE types | | | |
|---|---|---|---|---|
| Sly | LTR/Gypsy | DNA/DTM | LTR/unknown | Asterids |
| Vvi | LTR/Gypsy | LTR/Copia | LTR/unknown | |
| Car | LTR/Copia | DNA/DTC | DNA/Helitron | Eurosids I |
| Gma | LTR/Gypsy | LTR/unknown | DNA/DTM | |
| Cla | DNA/DTM | DNA/DTC | DNA/Helitron | |
| Csa | DNA/DTC | DNA/DTM | DNA/Helitron | |
| Fve | DNA/DTM | DNA/DTC | LTR/unknown | |
| Ath | DNA/Helitron | DNA/MuDR | LTR/Gypsy | Brassicaceae |
| Aly | DNA/Helitron | LTR/Gypsy | LTR/Copia | |
| Cru | DNA/Helitron | LTR/Gypsy | LTR/Copia | |
| Aal | DNA/Helitron | LTR/Gypsy | DNA/DTC | |
| Bol | DNA/Helitron | DNA/DTC | LTR/Copia | |
| Bna | DNA/Helitron | DNA/DTC | LTR/Gypsy | |
| Bju | DNA/Helitron | DNA/DTM | DNA/DTC | |
| Tha | LTR/Gypsy | LTR/Copia | LTR/unknown | |
| Slo | DNA/DTM | LTR/Copia | DNA/DTA | |
| Osa | LTR/Gypsy | DNAnona/MULE | DNAnona/Helitron | Monocots |
| Zma | LTR/Gypsy | LTR/Copia | DNA/CMC-EnSpm | |
| Atr | DNA/Helitron | LTR/unknown | LTR/Copia | |
| Afi | LTR/Copia | LTR/unknown | LTR/Gypsy | Non-flowering plants |
| Smo | LTR/Gypsy | DNA/DTM | LTR/unknown | |
| Mpo | DNA/DTM | LTR/Gypsy | DNA/DTC | |
| Ppa | LTR/Gypsy | LTR/unknown | long_terminal_repeat | |
| Cre | LTR/unknown | DNA/DTT | DNA/DTA | |
| Vca | LTR/Gypsy | DNA/DTM | DNA/DTC | |

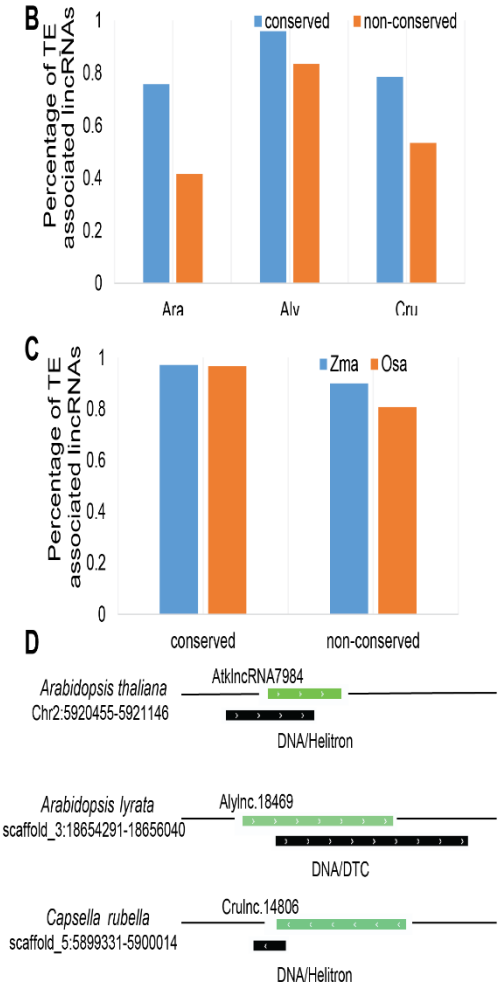



**Figure 3.28**: **Transposable elements (TEs) drive the evolutionary origins of lincRNAs.** (A) Top 3 TE types in terms of percentage of lincRNAs overlapping with a TE. (B) Percentage of non-conserved and conserved lincRNAs overlapping with TEs in *Arabidopsis thaliana (Ath), Arabidopsis lyrata (Aly), Capsella rubella (Cru).* (C) Percentage of non-conserved and conserved lincRNAs overlapping with TEs in *Oryza sativa* and *Zea mays*. (D) Schematic representation of lincRNAs in *Arabidopsis thaliana (Ath), Arabidopsis lyrata (Aly), Capsella rubella (Cru).* Green bars mean lincRNAs while the black bar is TEs.

### 3.2.8 LincRNAs in non-flowering plants

Based on direct comparison of lincRNA sequences from each of the three non-flowering plants (the model alga *Chlamydomonas reinhardtii*, the model land plants *Physcomitrella patens* and *Marchantia polymorpha*) to those from other plant species, 65 (12.8%), 232 (7.9%), 369 (7.0%) conserved lincRNAs were found in *C. reinhardtii*, *P. patens* and *M. polymorpha*, respectively. However, none of these sequence-based conserved lincRNAs could be detected by synteny-based approach, probably because of disruption of the syntenic blocks

during the long history of plant evolution. Nevertheless, the existence of sequence-based homologous lincRNAs in non-flowering plants suggests they may have potentially conserved function(s). Based on co-expression network analysis, several meristem development related modules (e.g. Mpo.Module3, 20, 32) were identified in *M. polymorpha*, and some of the lincRNAs of these three modules were highly expressed in reproductive tissues and sporophytes, and potentially linked with flower and anther development (**Figure 3.29C, Figure 3.30A**). These lincRNAs may represent the most ancient functional lincRNAs. Furthermore, conserved lincRNAs in *M. polymorpha* were linked with conserved non-coding sequences (CNSs) (**Figure 3.29A**), and had a higher phastCons score than the non-conserved ones (**Figure 3.29B**). Two conserved lincRNAs within the meristem-related modules were illustrated here (**Figure 3.30B, C**). In summary, these conserved lincRNAs in non-flowering plants make them the best candidates for the most conserved lincRNAs in the model flowering plants such as *Arabidopsis thaliana* and *Oryza sativa.*

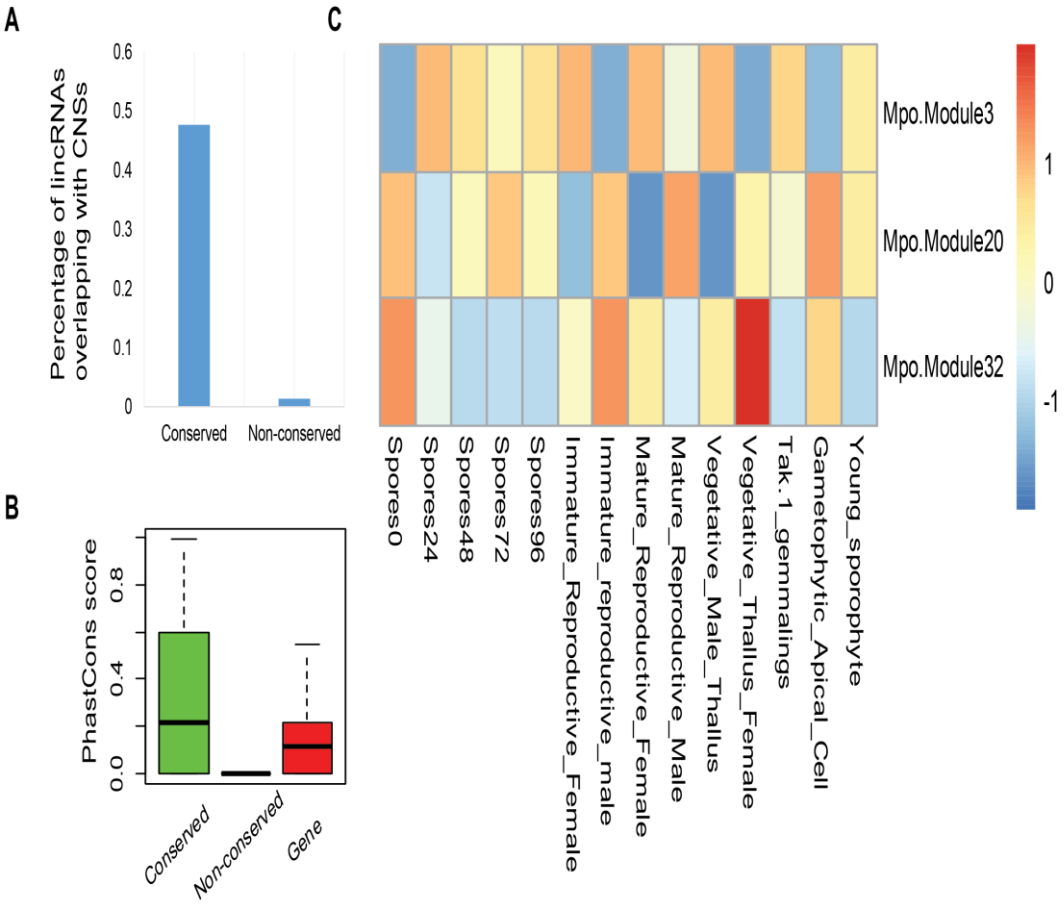



**Figure 3.29: Conserved lincRNAs in the land plant *Marchantia polymorpha*.** (A) Conserved lincRNAs in *Marchantia polymorpha* are enriched in conserved non-coding sequences (CNSs). (B)

Sequence conservation (26 plants genomes PhastCons scores) for conserved, non-conserved lincRNAs, PCGs in *Marchantia polymorpha*. (C) Expression pattern (eigengenes of each module) of meristem related modules in tissues of *Marchantia polymorpha*. Co-expression networks involving PCGs and lincRNAs in the land plant *Marchantia polymorpha* were constructed by WGCNA.

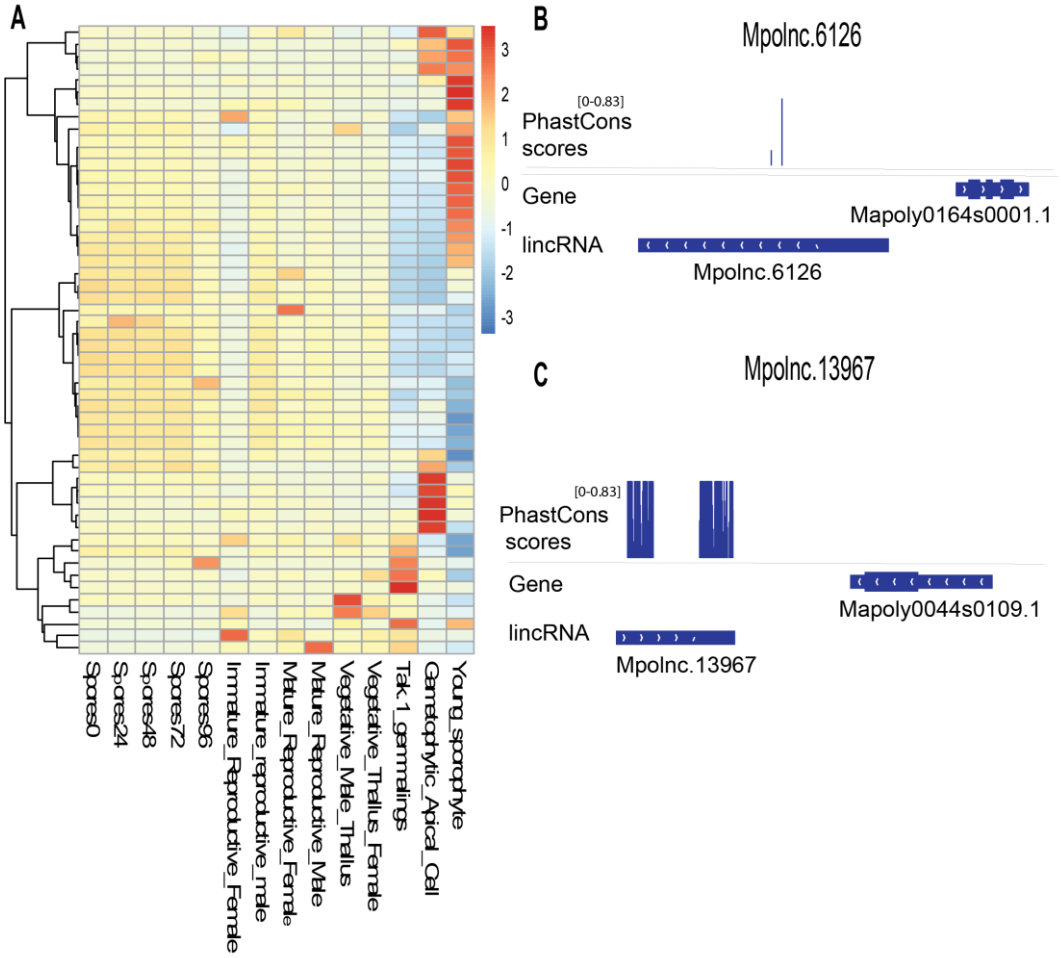



**Figure 3.30: LincRNAs in the land plant *Marchantia polymorpha*.** (A) The expression pattern across developmental tissues for conserved lincRNAs in meristem related modules (Mpo.Module3, 30, 32). (B) PhastCons scores for the lincRNA Mpolnc.6126 (Mpo.Module32). (C) PhastCons scores for the lincRNA Mpolnc.13967 (Mpo.Module3).

## 4. Discussion

We identified a comprehensive set of lincRNAs that show activity in reproductive and meristem tissues in *Arabidopsis thaliana*. Our data suggest a non-random spatial distribution of lincRNAs in the genome, allowing to classify the lincRNAs into TE-associated and non-TE-associated lincRNAs with different chromatin states in flowers. Chromatin characteristics and TF binding data suggest that lincRNA activity is regulated in a manner that is mechanistically different from that of protein-coding genes. LincRNAs are highly developmental stage-specific, associated with binding of master regulatory transcription factors. Moreover, we identified thousands of lincRNAs with the universe pipeline (Kapusta et al. 2014) in 26 plant species including non-flowering plants, and allow us to infer sequence conserved and synteny based homolog lincRNAs, and explore conserved characteristics of lincRNAs during plants evolution. Most lincRNAs evolve rapidly in terms of both sequence and expression pattern, which suggests higher transcriptional gain and loss of lincRNAs in plants.

### 4.1 Limitation of our study in identification of flower-related lincRNAs

After collecting a large number of flower/meristem polyA and total mRNA-seq datasets, thousands of flower-related lincRNAs were identified. In each dataset, ~100 lincRNAs could be detected by applying a polyA-RNA-seq analysis routine and after several filtering steps (e.g. transcript length, coding potential, and lowly expressed artifacts). However, in datasets from more specialized tissue types (e.g. laser capture microdissection (LCM), fluorescent activated cell sorting (FACS), and nuclear tagging in specific cell-types (INTACT)-sorted tissues), hundreds of lincRNAs could be identified (**Supplemental Table S1**). For example, 992 lincRNAs were identified in 1ld meristems of an INTACT dataset. Our results confirm that lincRNAs are typically lowly expressed and more tissue/stage specific than protein coding genes, which may limit or impair the identification of many lincRNAs in bulk tissue samples. Many lincRNAs such as circRNAs do not possess polyA tails (Di et al. 2014) and are thus not represented in poly(T) selected RNA-seq datasets. In our total RNA-seq datasets, several total RNA-seq specific lincRNAs could be identified, which suggests potential regulation by the RNA processing pathway. RACE-based identification of full-length lincRNAs (e.g. *APOLO*; Ariel et al. 2014; Ariel et al. 2020) suggested that RNA-seq typically identify only fragments of lincRNAs. Novel methodologies such as Iso-seq (Pacbio platform) could be used to obtain a more comprehensive overview of isoforms and full length lincRNAs (Lagarde et al. 2017). Therefore, a high-depth total RNA-seq library is preferentially chosen for the identification of lincRNAs because of missing splicing isoforms for lincRNAs by low depth of the library. When lncRNAs

were enriched and sequenced by *RACE* coupled with long-read high-throughput sequencing (RACE-seq), Lagarde et al found that lncRNAs are subject to alternative splicing like mRNAs (Lagarde et al. 2016).

We have noticed differences in the number of lincRNAs identified in our study with Liu et al, 2012 (**Figure S1A**). In our study, we mainly focus on the reproductive tissues and meristems listed in **Supplemental Table S1.** In the study of Liu et al, 2012, they performed RNA-seq on 4 tissue types in which we only used one tissue type (GSM946223, flowers) and not all of them. The cause of the observed differences is not only because of the high expression specificity and low overall abundance of lincRNAs but also due to the differences in the script and criteria that have been used in previous studies and this study to identify expressed transcripts.

**4.2 LincRNAs space in plant genomes is far from completeness**

Even though a large number of lincRNAs have been identified in diverse plants, thousands of new lincRNAs were still identified in this study by using publicly available RNA-seq datasets presumably due to their high tissue specificity and low expression levels. For example, only 12% of Ath lincRNAs and 26% of Gma lincRNAs identified in this study were overlapping with previously found (Szcześniak et al. 2016; Golicz et al. 2018). Inter-individual variations in the same plant species create a hinder for the comprehensive annotation of lincRNA in plants. This was not explored in this study but it has been shown in human that lincRNAs of primary granulocytes exhibited expressional viability and variability in different individuals (Kornienko et al. 2016). Furthermore, lincRNAs are heterogeneous groups of RNAs, present in different forms (e.g. linear or circular), possessing diverse properties (e.g. with polyA or without polyA tail), and showing variable stability (i.e. stable or unstable RNAs). CircRNAs are closed RNAs formed by back-joining of splicing acceptor and donor. CircRNAs and lincRNAs without polyA tails cannot be identified using the RNA-seq datasets used in this study. Some types of lincRNAs such as promoter upstream transcripts (PROMPTs) are unstable and degraded rapidly by nuclear RNA decay pathways, and thereby can only be seen in mutants of components in the exosome (Thieffry et al. 2020b). Finally, the current gene model of lincRNAs is inaccurate due to the limitation of Illumina RNA-seq and thereby molecular techniques such as RACE would be needed for verification. The third generation of sequencing technology such as Pacbio/SMRT and Nanopore can directly sequence the full length of lincRNAs, as demonstrated in studies of human lincRNAs (Lagarde et al. 2017).

**4.3 A subset of lincRNAs is associated with TEs**

After the identification of flower-related lincRNAs in Arabidopsis, we found that a subset

of them are associated with TEs. Many features, such as length, conservation, and histone modification status can be investigated to differentiate PCGs, TE-associated lincRNAs and non-TE-associated lincRNAs. This can help to define and classify lincRNAs and predict potential molecular functions (Quinn and Chang 2016). LincRNAs and PCGs fundamentally differ in many genomic features, such as chromatin state. Most TE-lincRNAs are associated with 24nt-siRNAs, suggesting that they constitute their precursor transcripts generated by Pol IV and Pol V activity, and thereby are components of the RdDM pathway silencing TEs epigenetically. TE-associated siRNAs have been implicated in disease resistance and the formation of hybridization barriers (Cho 2018). For example, TE-siR815, which is generated from a TE-associated lincRNA precursor in a *WRKY45* (LOC_Os05g25770) intron can cause methylation of its target gene *siR815 Target 1* (*ST1*, LOC_Os08g10150) through the RdDM pathway in rice (Zhang et al. 2016). Similarly, a retrotransposon-derived lincRNA named MIKKI could act as endogenous target mimics which has two mismatches at miRNA171 binding and cleavage sites and thereby blocks miRNA171 targeting *SCARECROW-Like* (*SCL*) transcription factor transcripts (Cho and Paszkowski 2017). Additionally, our and others' (Liu et al. 2015) results suggest that RNA pol II also contributes to the biogenesis of TE- and non-TE lincRNAs (**Figure 3.8A**). The question remains how many lincRNAs have a biological function, rather than just being by-products of pervasive RNA pol II activities. Our understanding of lincRNA functions in plants is only confined to a small number of well-known examples. Genome-wide prediction of potential lincRNA functions often uses the identification of neighboring genes (as the potential target gene in *cis*) and of co-expressed genes (in *trans*). Validation of lincRNA functions is technically challenging because it is important to distinguish them from the roles of the DNA sequence that encodes them (Bassett et al. 2014). For example, T-DNA mutants disrupt DNA sequence and may thereby impair lincRNA activity. However, it is not possible to distinguish phenotypic consequences caused by loss of lincRNA activity or by disruption of other regulatory functions of the DNA (e.g. TF DNA-binding). Furthermore, lncRNAs might have redundant, context-dependent, or minor functional roles as demonstrated by recent large-scale CRISPR deletion studies in zebrafish (Goudarzi et al. 2019) and C. *elegans* (Wei et al. 2019). Therefore, the classification of lincRNAs into different groups according to their association with genomic features, such as TEs, can help to distinguish potential functions and origins of this highly heterogeneous class of RNA molecules.

### 4.4 LincRNAs with potential roles in flower development

Using stage-specific transcriptome profiling in flower development, we found that many lincRNAs are more stage-specifically expressed than protein coding genes, in agreement with

previous findings (Liu et al. 2012). The dynamic expression of lincRNAs during the transition to flowering and flower morphogenesis indicates that these lincRNAs are developmentally regulated, possibly associated with DNA-binding of floral master regulators (e.g. AP1 and SEP3). Many lincRNA loci that are neighbors to flower developmental regulatory genes are dynamically expressed during floral transition and/or flower development. Thus, these lincRNAs represent candidates implicated in functions in flower development. For example, AtklncRNA5190, which is located in the neighborhood of the *AP1* locus, is specifically expressed in the shoot apical meristem at floral transition (2 days after transfer to inductive long-day conditions (You et al. 2017), whereas AtklncRNA19354 (near *AP2*) shows the highest expression one day later (You et al. 2017) (**Supplemental Table S1**). Examples from the animal field show that lncRNAs can have instructive roles during patterning processes in development. For example, it was demonstrated that lncRNAs are associated with homeotic Hox genes determining the body plan in animals, such as *HOTAIR* (Rinn et al. 2007; Amândio et al. 2016; Selleri et al. 2016). Since similar homeotic functions are exerted by MADS-domain TFs in plants, we postulate that lncRNAs could participate in Arabidopsis flower patterning either up- or downstream of the homeotic genes. The comprehensive set of floral lincRNAs identified here provides a valuable resource about candidate lincRNAs with potential roles in the flower development of *Arabidopsis*.

**4.5 LincRNAs are associated with TF DNA-binding in flower development**

Many lincRNAs are found to be closely associated with genomic regions bound by homeotic and other important developmental TFs. Because of their association with open chromatin and their location distal to protein-coding genes, these lincRNAs can represent enhancer RNAs (enhancer-associated lincRNAs). LincRNA expression dynamics are often correlated with that of their neighboring genes and in agreement with patterns of histone modifications (e.g. H3K4me3 and H3K27me3). Enhancer-associated lincRNAs have been described to function in target gene regulation through diverse mechanisms (in *cis* or *trans*). In *cis*, enhancer RNAs may induce and promote the formation of DNA loops between distal regions and promoters of PCGs (Ounzain and Pedrazzini 2015). In *trans*, enhancer RNAs activated by TFs could regulate remote PCGs by recruitment of histone modification enzymes and the transcription machinery (Ounzain and Pedrazzini 2015). For example, the lncRNA *RACER* was identified as a novel TF Tbx5-dependent lncRNA required for the regulation of the calcium-handling *Ryr2* gene expression in cardiac rhythm development and thereby the TF Tbx-dependent enhancer transcripts could be parts of the TF gene regulatory network (Yang et al. 2017). In *Arabidopsis*, an intronic lincRNA called *AG-incRNA4* from the second intron of

*AG* can recruit CURLY LEAF (CLF), an H3K27 trimethylation component of the Polycomb Repressive Complex (PRC) 2, to repress *AG* expression in leaves (Wu et al. 2018).

Enhancer activities are associated with stage-specific expression patterns in flower development (Yan et al. 2019). Previously, AP1 and SEP3 have been proposed to act as plant pioneer TFs by the opening of chromatin structures through recruitment of chromatin remodeling factors (Smaczniak et al. 2012; Yan et al. 2016). Our results showed that enhancer associated lincRNA expression is positively correlated with activities of neighboring protein coding genes, and may be utilized to distinguish activating and repressive activities of 'dual-function' TFs, or context-dependent TF activities (Azofeifa et al. 2018). It suggests that lincRNA associated enhancer activity could be used in a predictive manner to distinguish activating from repressive TF activities (Azofeifa et al. 2018). For example, we found that the *linc-AP2* locus spanning two enhancers is bound by multiple master regulator TFs including LFY, AP1, and SEP3. During flower development, binding of AP1 is associated with an increase in expression of the lincRNA, which consequently results in increased accessibility of the two corresponding enhancers, thereby contributing to the activation of *AP2*. On the other hand, strong overexpression of *LINC-AP2* in plants infected with Turnip crinkle virus (TCV) is associated with repression of *AP2* and contribute to abnormal structures of flowers in *Arabidopsis* (Gao et al. 2016), suggesting dosage-dependent effects of lincRNA levels in enhancer control. However, in normal conditions (WT), the expression of *LINC-AP2* and *AP2* are both increasing during early stages of flower morphogenesis, suggesting a joined activation. It will be interesting to analyze how different levels of *LINC-AP2* are mechanistically associated with activation vs. downregulation of *AP2* expression.

In summary, we propose a model: plant pioneer TFs such as AP1 and SEP3 bind to poorly accessible enhancer regions, induce enhancer associated lincRNA transcription which then assists in promoting an open chromatin status (**Figure 4.1**). Enhancer-associated lincRNAs precedes chromatin accessibility might be true only for a limited number of lincRNAs. There are 1038 enhancer associated lincRNAs in which only 337 (32.5%) lincRNAs display expression changes during flower development. Thus, our results show that flower related lincRNAs are associated with enhancer and might be components of the floral gene regulatory network.

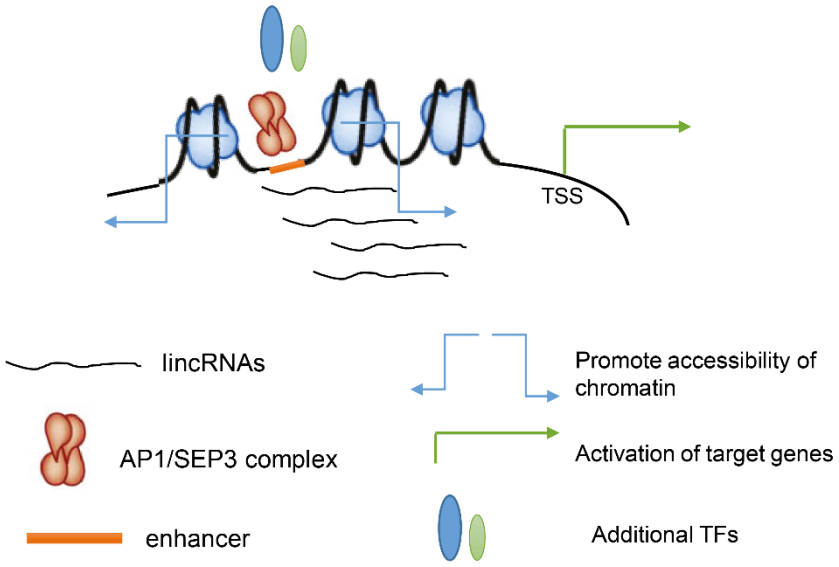



**Figure 4.1**: A working model for enhancer associated lincRNAs**.** Plant pioneer TFs such as AP1 and SEPs could bind into enhancers in closed chromatin and induce enhance associated lincRNAs transcription which then promotes open chromatin in enhancers and thereby activation of target genes.

### 4.6 Evolutionary landscape of lincRNAs across land plant species

The evolutionary landscape of lincRNAs has been explored in serval clades of eukaryotic species such as the *Brassicaceae* family (Mohammadin et al. 2015), monocots (Wang et al. 2015a), and vertebrates (Hezroni et al. 2015; Necsulea et al. 2014; Washietl et al. 2014; Sarropoulos et al. 2019; Bu et al. 2015). These studies found that, compared to PCGs, lincRNAs are shorter in length, have fewer exons, show lower expression levels and higher tissue-specificity. In addition, primary sequences of lincRNAs diverge faster than that of PCGs in both plants and animals. Therefore, evolutionary distance between species can significantly influence the number of identified homologous lincRNAs and the length of alignable sequence segments between homologous lincRNAs as demonstrated in this study. LincRNAs might be intermediate between neutrally evolving sequences and protein coding genes and their functionality may be conferred by their small conserved motifs. The conserved sequence patches within lincRNAs are potentially important for the functionality of lincRNAs. They provide binding sites for transcription factors and RNA binding proteins (RBPs) and also can be translated into small open reading frames (sORFs) (Ruiz-Orera et al. 2019). In this study, we found that binding sites of homeotic proteins such as AP1 and SEP3 were enriched in the promoters of conserved plant lincRNAs with a potential role in flower development. LincRNAs

usually show tissue-specific expression patterns, it is thus necessary to use same tissues for expression comparison. Rapid divergence of lincRNA sequences would alter or abolish the functionality of lincRNAs, but a study in zebrafish showed that homologous lincRNAs with poor sequence conservation could still retain their conserved functionality (Ulitsky et al. 2011), suggesting that the functionality of lincRNAs may depend on short sequence motifs (Ulitsky 2016) or their secondary structures. For example, the photomorphogenesis related lincRNA *HID* exhibits conserved sophisticated secondary structures between *Arabidopsis* and rice (Wang et al. 2014b).

**4.7 Transposable elements play important roles in the origin of plant lincRNAs**

Gain and loss of lincRNA in the evolution history of plants are faster than that of PCGs (Ulitsky 2016). In this study, we used diverse plant species and RNA-seq datasets in identification of lincRNAs. Despite differences in the quality of genomes and RNA-seq datasets, which made it difficult to estimate and compare the exact number of lincRNAs in each plant genome; however, it seems that the number of lincRNAs is positively correlated with the size of genomes, particularly in plant genomes with high propotion of TEs. We hypothesize that this may be partially explained by diverse contributions of TEs in the origin of lincRNAs. TEs might contribute to the exonization of lincRNAs just like cases in mRNAs (Sela et al. 2010), provide transcription start sites, splice sites and polyA sites (Kapusta et al. 2013), Subsequently, these TE-derived elements or motifs became the sources of functional elements of lincRNAs (Lee et al. 2019; Johnson and Guigó 2014). We found that a significant number of lincRNAs are associated with TEs in all plant species investigated in this study. Some of these TE-associated lincRNAs could actually be direct transcription products of TEs. In most plant families (except Brassicaceae), the top type of TE associated with lincRNAs was retrotransposons, consistent with the previous finding that ancestral TEs play important roles in the origination of lincRNAs (Wang et al. 2015d). In plants, TE-associated lincRNAs could be induced by abiotic stresses such as salt and cold treatments (Wang et al. 2017a). In humans, TEs drive tissue-specific expression in stem cells and thus shape the function and evolution of lincRNAs (Kelley and Rinn 2012). Furthermore, sequence similarity between some lincRNAs in different species often overlap with conserved enhancer elements which drive expression of target genes (Hezroni et al. 2017).

**4.8 Comparative genomics approaches to understand lincRNAs**

Comparative genomic approaches are powerful tools to infer and prioritize the potential functions of genes and molecular mechanisms (mode of action) as demonstrated in functional studies of PCGs and miRNAs. For example, some miRNAs such as miR156 and miR159 are

highly conserved in plants, including non-flowering plants. Comparative genomic analyses have facilitated the identification and functional characterization of the conserved miRNAs in different plant species (Zhang et al. 2006). This principle should also be applicabe for lincRNAs. Identification and functional characterization of lincRNAs in model species such as *A. thaliana* would give opportunities to understand their homologous lincRNAs in non-model organisms that do not have well defined molecular and genetic tools. Several approaches have been used to identify homologous lincRNAs in plants. One is whole genome alignment. This has been widely used in the animal field because it is available in the public databases such as UCSC genome browse. However, many potential homologous lincRNAs could be missed out when using this approach as lincRNA homology quite often can only be found in short sequence patches, it is therefore critical to find a suitable cut-off value when applying this approach otherwise the power of this method would be compromised. Another is to direct compare sequences using alignment tool such as blast. This approach is computationally more efficient than the approach of whole genome alignment. In addition, based on syntenic relationship of neighboring PCGs, positional conservation can also be used to identify homologous lincRNAs. However, if the intergenic region of interest contains multiple lincRNAs, addition information would be required to determine the authentic homologous lincRNAs. Conservation at both sequence and syntenic position would strongly suggest homologous relationship but the number of such lincRNAs are very small, presumably due to rapid sequence divergence and/or disruption of syntheny by multiple rounds of whole genome duplication and other forces of genome rearrangement.

The lincRNAs identified in this study, particularly the conserved ones, provide resources for identification of their homologs in many newly sequenced plant genomes. Additionally, many excellent algorithms have been developed for better aligning and comparing lincRNA sequences, which would enhance sequence homology based lincRNA identification(Chen et al. 2016; Lin et al. 2011; Washietl et al. 2011). Non-synonymous to synonymous changes (dN/dS) is often used to evaluate evolutionary constraints on PCGs but its application in lincRNAs is still absent.

## 5. Future perspectives

Mounting evidence shows the involvement of lincRNAs in wide ranges of biological processes, including development and stress responses. LincRNAs act in *cis* or in *trans* to regulate the function of their target genes through diverse mechanisms that involve interactions with DNA, RNA, and proteins. However, despite the vastness of the diversity of lincRNAs molecular mechanisms and functions, our understanding of most plant lincRNAs is still elusive and unclear. There are at least a couple of reasons. Firstly, the effects of lincRNAs might only be observed under specific conditions given that the expression of most lincRNAs is highly tissue/condition-specific. Secondly, lincRNAs represent a heterogeneous group of RNA molecules in plants. Several subclasses of lincRNAs (e.g. enhancer RNAs) are largely coupled with regulatory DNA sequences (e.g. TFBSs), which makes it difficult to assess their specific functions. The development of technologies is imperative to understand the molecular mechanisms of lncRNAs (Ariel et al. 2020).

In our study, we demonstrate lincRNAs are associated with enhancers in *Arabidopsis* and these enhancer-associated lincRNAs are components of the floral gene regulatory network. Besides, one of the examples was selected to validate the model. In order to obtain more lincRNAs regulators of flower development, efficient computational methods are urgently needed to predict functional lncRNAs for experimental validation among large numbers of lincRNAs identified. Furthermore, large scale functional screens of lincRNAs by CRISPR/Cas9-based mutagenesis has been established in human and flies, although only a small percentage of lincRNAs showed context-specific phenotypic changes (Liu et al. 2017). A similar system has yet to be developed for plant lincRNAs although large-scale mutagenesis populations have been created in several plant species by the transformation of sgRNA libraries targeting protein-coding genes (Liu et al. 2020; Jacobs et al. 2017; Bai et al. 2020; Zhang et al. 2020; Lu et al. 2017; Meng et al. 2017). Therefore, it should be feasible to practice a large-scale genome-wide screen of a reduced number of computationally selected candidate lincRNAs so that it can largely decrease the costs of designing gRNAs. Additionally, the methods for designing sgRNAs or pgRNAs (pair guide RNAs) targeting lincRNAs can also be considered.

For example, in this study, we identified Ara.lnc19175/*linc-AP2* with dynamic expression during flower development and it is associated with two putative enhancers within the body regions of the lincRNA itself, which are bound by multiple master TF regulators such as AP1 and SEP3. Previously, *linc-AP2* was found to be upregulated in Turnip crinkle virus (TCV)-infected conditions while the A-class gene *AP2* was downregulated. Additionally, the overexpression lines of Ara.lnc19175 in *Arabidopsis* displayed abnormal floral structures (e.g. shorter stamen filaments) in TCV-infected conditions(Gao et al. 2016). In our system, both the

expression of *linc-AP2* and *AP2* are increasing at the same time when AP1 is activated/induced by GR. We hypothesize that the activation is largely caused by enhancers in the lincRNA region and the functionality of the enhancers is related to the lincRNA. Further experiments need to be designed to verify this inconsistency. One thing that has to be pointed out is the entangling effects between the cis-regulatory elements (enhancers) inside the lincRNA, the lincRNA transcript itself, and even the lincRNA transcription/alternative splicing (Gil and Ulitsky 2020; Engreitz et al. 2016). Therefore, multiple perturbation approaches (e.g. overexpression and CRISPR/cas9) are necessary to unlink *linc-AP2* transcripts with its associated TFBSs there in order to understand the molecular mechanisms of *linc-AP2* regulating the neighboring gene *AP2*. Studies from the human field provided several excellent examples to dissect linked effects between the lincRNA and the coupled cis-regulatory elements inside, such as *Lockd* (lncRNA near *Cdkn1b*) (Paralkar et al. 2016) and *Haunt* (Yin et al. 2015). In addition to the in *cis* targets of *linc-AP2*, the in *trans* targets cannot be overlooked. It is worthwhile to express *linc-AP2* from a different genomic context to check whether the *cis* activity could be restored after transfection. If this *trans* complementation could not rescue the phenotypes (e.g. flower development defects and *AP2* expression), the *cis*-regulatory activity can be verified. Both our study and other studies provide clues that neighboring genes of lincRNAs are enriched in TFs (Gil and Ulitsky 2020). Furthermore, there are also regulatory elements bound by AP2 potentially modulating the *linc-AP2* itself and thus it established auto-regulatory loops during flower development. It provides an excellent model to understand the *cis* regulatory mechanism of the lincRNA. Finally, *linc-AP2* is also overlapping with one TE. However, the functional significance of this TE within *linc-AP2* is quite unknown. The *linc-AP2* spanning two contrasting chromatin states (euchromatin and heterochromatin) might provide a unique model to understand the chromatin state transition during flower development.

LincRNAs (long intergenic non-coding RNAs) are not the only types of lncRNAs demonstrated to have important functions in floral transition and flower development. Other types including antisense lncRNAs and intronic lncRNAs also need to be considered. For example, many lncRNAs (e.g. *Ef-cd, MAS,* and *COLDAIR*) are antisense to the important MADS genes such as *SOC1*, *MAFs,* and *FLC*). We wonder whether these antisense lncRNAs residing in the MADS TF family are widespread phenomena not only in *Arabidopsis thaliana* but also in the whole plant family. Furthermore, the next question is what functions they have if it is the case. Additionally, the category of intronic lncRNAs also has several examples (e.g. *COOLAIR* and *AG-incRNA4*) in the intron of the MADS domain TF family. What about other MADS family members in *Arabidopsis thaliana*? What about other plants? Is it widespread? How spread are

they?

We sequenced both polyA RNA-seq and total RNA-seq in the floral induction system and this allows us to understand other types of lncRNAs: lncRNAs without polyA. One of the examples among these is the circRNA *circSEP3* which impacts the expression of SEP3 and has an important function in flower morphogenesis (Conn et al. 2017). The regulatory mechanism of these kinds of lncRNAs is promising to be understood with the power of large numbers of high-throughput datasets such as ChIP-seq and DNase-seq.

The genome-wide identification of lincRNAs across the whole plant lineages and most lincRNAs are species-specific. There are still some conserved lincRNAs to some extent and they are under evolutionary constraint. Therefore, we wonder what functions of these conserved lincRNAs they have and these candidates need to be investigated in the future.

Finally, we need to investigate how we can effectively utilize the knowledge of beneficial lncRNAs in breeding programs to develop novel plant germplasm and elite crop varieties. An excellent example of this is provided by *Ef-cd* that promotes early maturity without yield penalty probably due to better nitrogen utilization and photosynthesis in rice. It functions as a dominant gene as plants homozygous or heterozygous for *Ef-cd* showed better agronomic performance compared to plants without *Ef-cd*. It thus is valuable for rice breeding. Fang et al. (2019) has developed molecular markers completely linked with *Ef-cd*, which can be used to identify new early maturity rice germplasm containing *Ef-cd* and to introgress *Ef-cd* into elite rice cultivars to further improve their maturity and agronomic performance based on marker-assisted selection. For *LDMAR* and *PMS1T*, base editing can be used to change the unfavorable alleles into favorable ones as single nucleotide polymorphisms seem to be the cause for changes in fertility. These examples show that utilizing knowledge on plant lncRNA functions can open new possibilities for plant breeding research, thereby improving crop quality and performance.

## References


Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KMM, Cao J, Chae E, Dezwaan TMM, Ding W, et al. 2016. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* **166**: 481–491.

Amândio AR, Necsulea A, Joye E, Mascrez B, Duboule D. 2016. Hotair Is Dispensible for Mouse Development ed. G.S. Barsh. *PLOS Genet* **12**: e1006232.

Ariel F, Jegu T, Latrasse D, Romero-Barrios N, Christ A, Benhamed M, Crespi M. 2014. Noncoding transcription by alternative RNA polymerases dynamically regulates an auxin-driven chromatin loop. *Mol Cell* **55**: 383–396.

Ariel F, Lucero L, Christ A, Liu C, Benhamed M, Crespi M, Mammarella MF, Jegu T, Veluchamy A, Mariappan K, et al. 2020. R-Loop Mediated trans Action of the APOLO Long Noncoding RNA. *Mol Cell* **77**: 1–11.

Ariel F, Romero-Barrios N, Jégu T, Benhamed M, Crespi M. 2015. Battles and hijacks: Noncoding transcription in plants. *Trends Plant Sci* **20**: 362–371.

Azofeifa JG, Allen MA, Hendrix JR, Read T, Rubin JD, Dowell RD. 2018. Enhancer RNA profiling predicts transcription factor activity. *Genome Res* **28**: 334–344.

Bai M, Yuan J, Kuang H, Gong P, Li S, Zhang Z, Liu B, Sun J, Yang M, Yang L, et al. 2020. Generation of a multiplex mutagenesis population via pooled CRISPR-Cas9 in soya bean. *Plant Biotechnol J* **18**: 721–731.

Bardou F, Ariel F, Simpson CG, Romero-Barrios N, Laporte P, Balzergue S, Brown JWS, Crespi M. 2014. Long Noncoding RNA Modulates Alternative Splicing Regulators in Arabidopsis. *Dev Cell* **30**.

Bassett AR, Akhtar A, Barlow DP, Bird AP, Brockdorff N, Duboule D, Ephrussi A, Ferguson-Smith AC, Gingeras TR, Haerty W, et al. 2014. Considerations when investigating lncRNA function in vivo. *Elife* **3**: 1–14.

Bazin J, Baerenfaller K, Gosai SJ, Gregory BD, Crespi M, Bailey-Serres J. 2017. Global analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation. *Proc Natl Acad Sci U S A* **114**: E10018–E10027.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 4–8.

Bu DC, Luo HT, Jiao F, Fang SS, Tan CF, Liu ZY, Zhao Y. 2015. Evolutionary annotation of conserved long non-coding RNAs in major mammalian species. *Sci China Life Sci* **58**: 787–798.

Budak H, Kaya SB, Cagirici HB. 2020. Long Non-coding RNA in Plants in the Era of Reference Sequences. *Front Plant Sci* **11**: 276.

Campalans A, Kondorosi A, Crespi M. 2004. Enod40, a Short Open Reading Frame-Containing mRNA, Induces Cytoplasmic Localization of a Nuclear RNA Binding Protein in Medicago truncatula. *Plant Cell* **16**: 1047–1059.

Cao M, Zhao J, Hu G. 2019. Genome-wide methods for investigating long noncoding RNAs. *Biomed Pharmacother* **111**: 395–401.

Carlevaro-fita J, Johnson R. 2019. Global Positioning System : Understanding Long Noncoding RNAs through Subcellular Localization. *Mol Cell* **73**: 869–883.

Carlevaro-Fita J, Polidori T, Das M, Navarro C, Zoller TI, Johnson R. 2019. Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. *Genome Res* **29**: 208–222.

Castaings L, Bergonzi S, Albani MC, Kemi U, Savolainen O, Coupland G. 2014. Evolutionary conservation of cold-induced antisense RNAs of FLOWERING LOCUS C in Arabidopsis thaliana perennial relatives. *Nat Commun* **5**: 4457.

Chekanova JA. 2015. Long non-coding RNAs and their functions in plants. *Curr Opin Plant Biol* **27**: 207–216.

Chekanova JA, Gregory BD, Reverdatto S V., Chen H, Kumar R, Hooker T, Yazaki J, Li P, Skiba N, Peng Q, et al. 2007. Genome-Wide High-Resolution Mapping of Exosome Substrates Reveals Hidden Features in the Arabidopsis Transcriptome. *Cell* **131**: 1340–1353.

Chen D, Fu L-Y, Zhang P, Chen M, Kaufmann K. 2019. ChIP-Hub: an Integrative Platform for Exploring Plant Regulome. *bioRxiv* 768903.

Chen D, Yan W, Fu LY, Kaufmann K. 2018. Architecture of gene regulatory networks controlling flower development in Arabidopsis thaliana. *Nat Commun* **9**: 1–13.

Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, et al. 2020. Pervasive functional translation of noncanonical human open reading frames. *Science (80- )* **367**: 1140–1146.

Chen J, Shishkin AA, Zhu X, Kadri S, Maza I, Guttman M, Hanna JH, Regev A, Garber M. 2016. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biol* **17**: 1–17.

Chen LL. 2016. The biogenesis and emerging roles of circular RNAs. *Nat Rev Mol Cell Biol* **17**: 205–211.

Cho J. 2018. Transposon-Derived Non-coding RNAs and Their Function in Plants. *Front Plant Sci* **9**: 1–6.

Cho J, Paszkowski J. 2017. Regulation of rice root development by a retrotransposon acting as a microRNA sponge. *Elife* **6**.

Chu C, Quinn J, Chang HY. 2012. Chromatin isolation by RNA purification (ChIRP). *J Vis Exp* 3912.

Conn VM, Hugouvieux V, Nayak A, Conos SA, Capovilla G, Cildir G, Jourdain A, Tergaonkar V, Schmid M, Zubieta C, et al. 2017. A circRNA from SEPALLATA3 regulates splicing of its cognate mRNA through R-loop formation. *Nat Plants* **17053**: 4–8.

Crespi MD, Jurkevitch E, Poiret M, d'Aubenton-Carafa Y, Petrovics G, Kondorosi E, Kondorosi A, Crespil MD, Jurkevitchl E, Poiret M, et al. 1994. enod40, a Gene Expressed During Nodule Organogenesis, Codes for a Non-Translatable RNA Involved in Plant Growth. *EMBO J* **13**: 5099–5112.

Csorba T, Questa JI, Sun Q, Dean C. 2014. Antisense COOLAIR mediates the coordinated switching of chromatin states at FLC during vernalization. *Proc Natl Acad Sci U S A* **111**: 16160–16165.

Cui J, Jiang N, Meng J, Yang G, Liu W, Zhou X, Ma N, Hou X, Luan Y. 2019. LncRNA33732-respiratory burst oxidase module associated with WRKY1 in tomato- Phytophthora infestans interactions. *Plant J* **97**: 933–946.

Cui J, Luan Y, Jiang N, Bao H, Meng J. 2017. Comparative transcriptome analysis between resistant and susceptible tomato allows the identification of lncRNA16397 conferring resistance to Phytophthora infestans by co-expressing glutaredoxin. *Plant J* **89**: 577–589.

Datta R, Paul S. 2019. Long non-coding RNAs: Fine-tuning the developmental responses in plants. *J Biosci* **44**: 1–11.

de Velde J Van, Heyndrickx KS, Vandepoele K. 2014. Inference of transcriptional networks in Arabidopsis through conserved noncoding sequence analysis. *Plant Cell* **26**: 2729–2745.

Deforges J, Reis RS, Jacquet P, Sheppard S, Gadekar VP, Hart-Smith G, Tanzer A, Hofacker IL, Iseli C, Xenarios I, et al. 2019. Control of cognate sense mrna translation by cis-natural antisense RNAs. *Plant Physiol* **180**: 305–322.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* **22**: 1775–1789.

Di C, Yuan J, Wu Y, Li J, Lin H, Hu L, Zhang T, Qi Y, Gerstein MB, Guo Y, et al. 2014. Characterization of stress-responsive lncRNAs in Arabidopsis thaliana by integrating expression, epigenetic and structural features. *Plant J* **80**: 848–861.

Ding J, Lu Q, Ouyang Y, Mao H, Zhang P, Yao J, Xu C, Li X, Xiao J, Zhang Q. 2012a. A long

noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. *Proc Natl Acad Sci U S A* **317784**: 2654–2659.

Ding J, Shen J, Mao H, Xie W, Li X, Zhang Q. 2012b. RNA-Directed DNA Methylation Is Involved in Regulating Photoperiod-Sensitive Male Sterility in Rice. *Mol Plant* **5**: 1210–1216.

Drechsel G, Kahles A, Kesarwani AK, Stauffer E, Behr J, Drewe P, Rätsch G, Wachter A. 2013. Nonsense-Mediated Decay of Alternative Precursor mRNA Splicing Variants Is a Major Determinant of the Arabidopsis Steady State Transcriptome. *Plant Cell* **25**: 3726–3742.

Du Q, Wang K, Zou C, Xu C, Li WX. 2018. The PILNCR1 -miR399 Regulatory Module Is Important for Low Phosphate Tolerance in Maize 1. *Plant Physiol* **177**: 1743–1753.

Engreitz JM, Haines JE, Perez EM, Munson G, Chen J, Kane M, McDonel PE, Guttman M, Lander ES. 2016. Local Regulation of Gene Expression by lncRNA Promoters, Transcription and Splicing. *Nature* **539**: 452–455.

Engreitz JM, Sirokman K, McDonel P, Shishkin AA, Surka C, Russell P, Grossman SR, Chow AY, Guttman M, Lander ES. 2014. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. *Cell* **159**: 188–199.

Ernst J, Kellis M. 2017. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* **12**: 2478–2492.

Fabbri M, Girnita L, Varani G, Calin GA. 2019. Decrypting noncoding RNA interactions, structures, and functional networks. *Genome Res* **29**: 1377–1388.

Fan Y, Yang J, Mathioni SM, Yu J, Shen J, Yang X, Wang L, Zhang QQ, Cai Z, Xu C, et al. 2016. PMS1T, producing phased small-interfering RNAs, regulates photoperiod-sensitive male sterility in rice. *Proc Natl Acad Sci U S A* **113**: 15144–15149.

Fang J, Zhang F, Wang H, Wang W, Zhao F, Li Z, Sun C, Chen FF, Xu F, Chang S, et al. 2019. Ef-cd locus shortens rice maturity duration without yield penalty. *Proc Natl Acad Sci U S A* **116**: 18717–18722.

Fatica A, Bozzoni I, Rna N. 2014. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet Vol* **15**: 7–21.

Favorov A, Mularoni L, Cope LM, Medvedeva Y, Mironov AA, Makeev VJ, Wheelan SJ. 2012. Exploring Massive, Genome Scale Datasets with the GenometriCorr Package ed. H. Lapp. *PLoS Comput Biol* **8**: e1002529.

Fedak H, Palusinska M, Krzyczmonik K, Brzezniak L, Yatusevich R. 2016. Control of seed dormancy in Arabidopsis by a cis-acting noncoding antisense transcript. *Proc Natl Acad Sci U S A* **113**: E7846–E7855.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-vegas A, et al. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* **44**: 279–285.

Franco-Zorrilla JM, Valli A, Todesco M, Mateos I, Puga MI, Rubio-Somoza I, Leyva A, Weigel D, García JA, Paz-Ares J. 2007. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat Genet* **39**: 1033–1037.

Frazee AC, Pertea G, Jaffe AE, Langmead B, Salzberg SL, Leek JT. 2015. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat Biotechnol* **33**: 243–246.

Gagliardi D, Cambiagno DA, Arce AL, Tomassi AH, Giacomelli JI, Ariel FD, Manavella PA. 2019. Dynamic regulation of chromatin topology and transcription by inverted repeat-derived small RNAs in sunflower. *Proc Natl Acad Sci U S A* **116**: 17578–17583.

Gai YP, Yuan SS, Zhao YN, Zhao HN, Zhang HL, Ji XL, Balmer DA. 2018. A Novel LncRNA , MuLnc1 , Associated With Environmental Stress in Mulberry ( Morus multicaulis ). *Front Plant Sci* **9**: 1–13.

Gao R, Liu P, Irwanto N, Loh DR, Wong SM. 2016. Upregulation of LINC-AP2 is negatively correlated with AP2 gene expression with Turnip crinkle virus infection in Arabidopsis thaliana. *Plant Cell Rep* **35**: 2257–2267.

Giacomelli JI, Weigel D, Chan RL, Manavella PA. 2012. Role of recently evolved miRNA regulation of sunflower HaWRKY6 in response to temperature damage. *New Phytol* **195**: 766–773.

Gil N, Ulitsky I. 2020. Regulation of gene expression by cis-acting long non-coding RNAs. *Nat Rev Genet* **21**: 102–117.

Gil N, Ulitsky I, Gil N, Ulitsky I. 2018. Production of Spliced Long Noncoding RNAs Specifies Regions with Increased Enhancer Activity Report Production of Spliced Long Noncoding RNAs Specifies Regions with Increased Enhancer Activity. *Cell Syst* **7**: 537-547.e3.

Golicz AA, Singh MB, Bhalla PL. 2018. The long intergenic noncoding RNA (LincRNA) Landscape of the soybean genome. *Plant Physiol* **176**: 2133–2147.

Goudarzi M, Berg K, Pieper LM, Schier AF. 2019. Individual long non-coding RNAs have no overt functions in zebrafish embryogenesis, viability and fertility. *Elife* **8**.

Gowthaman U, García-Pichardo D, Jin Y, Schwarz I, Marquardt S. 2020. DNA Processing in the Context of Noncoding Transcription. *Trends Biochem Sci* **45**: 1009–1021.

Grzechnik P, Tan-Wong SM, Proudfoot NJ. 2014. Terminate and make a loop: Regulation of transcriptional directionality. *Trends Biochem Sci* **39**: 319–327.

Guo G, Liu X, Sun F, Cao J, Huo N, Wuda B, Xin M, Hu Z, Du J, Xia R, et al. 2018. Wheat miR9678 Affects Seed Germination by Generating Phased siRNAs and Modulating Abscisic Acid/Gibberellin Signaling. *Plant Cell* **30**: 796–814.

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet* **45**: 891–898.

Hawkes EJ, Hennelly SP, Novikova I V, Irwin JA, Dean C, Sanbonmatsu KY. 2016. COOLAIR Antisense RNAs Form Evolutionarily Conserved Elaborate Secondary Structures. *Cell Rep* **16**: 3087–3096.

Henriques R, Wang H, Liu J, Boix M, Huang LF, Chua NH. 2017. The antiphasic regulatory module comprising CDF5 and its antisense RNA FLORE links the circadian clock to photoperiodic flowering. *New Phytol* **216**.

Heo JB, Sung S. 2011. Vernalization-Mediated Epigenetic Silencing by a Long Intronic Noncoding RNA. *Science (80- )* **331**: 76–79.

Hetzel J, Duttke SH, Benner C, Chory J. 2016. Nascent RNA sequencing reveals distinct features in plant transcription. *Proc Natl Acad Sci U S A* **113**: 1–6.

Hezroni H, Ben-Tov Perry R, Meir Z, Housman G, Lubelsky Y, Ulitsky I, Perry RB, Meir Z, Housman G, Lubelsky Y, et al. 2017. A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biol* **18**: 1–15.

Hezroni H, Koppstein D, Bartel DP, Ulitsky I, Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. 2015. Principles of Long Noncoding RNA Evolution Derived From Direct Comparison of Transcriptomes in 17 Species. *Cell Rep* **11**: 1110–1122.

Hisanaga T, Okahashi K, Yamaoka S, Kajiwara T, Nishihama R, Shimamura M, Yamato KT, Bowman JL, Kohchi T, Nakajima K. 2019. A cis-acting bidirectional transcription switch controls sexual dimorphism in the liverwort. *EMBO J* **38**: 1–12.

Hsu PY, Calviello L, Wu HYL, Li FW, Rothfels CJ, Ohler U, Benfey PN. 2016. Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proc Natl Acad Sci U S A* **113**: E7126–E7135.

Hu L, Xu Z, Hu B, Lu ZJ. 2017. COME: A robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. *Nucleic Acids Res* **45**: 1–13.

Huang D, Feurtado JA, Smith MA, Flatman LK, Koh C, Cutler AJ. 2017. Long noncoding miRNA gene represses wheat β -diketone waxes. *Proc Natl Acad Sci U S A* **114**: E3149–E3158.

Huang L. 2018. Systematic identification of long non-coding RNAs during pollen development and fertilization in Brassica rapa. *Plant J* **1**: 203–222.

Hupalo D, Kern AD. 2013. Conservation and functional element discovery in 20 angiosperm plant genomes. *Mol Biol Evol* **30**: 1729–1744.

Hüttenhofer A, Schattner P, Polacek N. 2005. Non-coding RNAs: Hope or hype? *Trends Genet* **21**: 289–297.

Illustrations P. 2017. Plant pictures. https://figshare.com/collections/Plant_pictures/3701239/5.

Jabnoune M, Secco D, Lecampion C, Robaglia C, Shu Q, Poirier Y. 2013. A Rice cis-Natural Antisense RNA Acts as a Translational Enhancer for Its Cognate mRNA and Contributes to Phosphate Homeostasis and Plant Fitness. *Plant Cell* **25**.

Jacobs TB, Zhang N, Patel D, Martin GB. 2017. Generation of a collection of mutant tomato lines using pooled CRISPR libraries. *Plant Physiol* **174**: 2023–2037.

Jiang N, Cui J, Shi Y, Yang G, Zhou X, Hou X, Meng J, Luan Y. 2019. Tomato lncRNA23468 functions as a competing endogenous RNA to modulate NBS-LRR genes by decoying miR482b in the tomato -Phytophthora infestans interaction. *Hortic Res* **6**: 28.

Johnson R, Guigó R. 2014. The RIDL hypothesis : transposable elements as functional domains of long noncoding RNAs. *RNA* **20**: 959–976.

Kang Y, Yang D, Kong L, Hou M, Meng Y, Wei L, Gao G. 2017. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res* **45**: 12–16.

Kapusta A, Feschotte C, Review F, Kapusta A, Feschotte C. 2014. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet* **30**: 439–452.

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. 2013. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genet* **9**: e1003470.

Kelley D, Rinn J. 2012. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol* **13**: R107.

Keniry A, Oxley D, Monnier P, Kyba M, Dandolo L, Smits G, Reik W. 2012. The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r. *Nat Cell Biol* **14**: 659–665.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: A fast spliced aligner with low memory requirements. *Nat Methods* **12**: 357–360.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907–915.

Kim D, Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* **11**: 1650–1667.

Kim D, Sung S. 2018. Vernalization-triggered intragenic chromatin-loop formation by long noncoding RNAs. *Dev Cell* **40**: 302–312.

Kim DH, Xi Y, Sung S. 2017. Modular function of long noncoding RNA, COLDAIR, in the vernalization response. *PLoS Genet* **13**: 1–18.

Kindgren P, Ivanov M, Marquardt S. 2018. Transcriptional read-through of the long non-coding RNA SVALKA governs plant cold acclimation. *Nat Commun* **9**: 4561.

Kirn SH, Koroleva OA, Lewandowska D, Pendle AF, Clark GP, Simpson CG, Shaw PJ, Brown JWSS, Kim SH, Koroleva OA, et al. 2009. Aberrant mRNA Transcripts and the Nonsense-Mediated Decay Proteins UPF2 and UPF3 Are Enriched in the Arabidopsis Nucleolus. *Plant Cell* **21**: 2045–2057.

Klepikova A V., Logacheva MD, Dmitriev SE, Penin AA. 2015. RNA-seq analysis of an apical meristem time series reveals a critical point in Arabidopsis thaliana flower initiation. *BMC Genomics* **16**: 466.

Kopp F, Mendell JT. 2018. Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell* **172**: 393–407.

Kornienko AE, Dotter CP, Guenzl PM, Gisslinger H, Gisslinger B, Cleary C, Kralovics R, Pauler FM, Barlow DP. 2016. Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol* **17**: 14.

Krizek BA, Fletcher JC. 2005. Molecular mechanisms of flower development: An armchair guide. *Nat Rev Genet* **6**: 688–698.

Kryuchkova-mostacci N, Robinson-rechavi M. 2017. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* **18**: 205–214.

Kung JTY, Colognori D, Lee JT. 2013. Long noncoding RNAs: Past, present, and future. *Genetics* **193**: 651–669.

Kurihara Y, Matsui A, Hanada K, Kawashima M, Ishida J, Morosawa T, Tanaka M, Kaminuma E, Mochizuki Y, Matsushima A, et al. 2009. Genome-wide suppression of aberrant mRNA-like noncoding RNAs by NMD in Arabidopsis. *Proc Natl Acad Sci U S A* **106**: 2453–2458.

Kurihara Y, Schmitz RJ, Nery JR, Schultz MD, Okubo-Kurihara E, Morosawa T, Tanaka M, Toyoda T, Seki M, Ecker JR. 2012. Surveillance of 3′ Noncoding Transcripts Requires FIERY1 and XRN3 in Arabidopsis. *G3 Genes, Genomes, Genet* **2**: 487–498.

Lagarde J, Uszczynska-Ratajczak B, Carbonell S, Pérez-Lluch S, Abad A, Davis C, Gingeras TR, Frankish A, Harrow J, Guigo R, et al. 2017. High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat Genet* **49**: 1731–1740.

Lagarde J, Uszczynska-Ratajczak B, Santoyo-Lopez J, Gonzalez JM, Tapanari E, Mudge JM, Steward CA, Wilming L, Tanzer A, Howald C, et al. 2016. Extension of human lncRNA transcripts by RACE coupled with long-read high-throughput sequencing (RACE-Seq). *Nat Commun* **7**: 1–11.

Lander ES. 2014. Ribosome profiling provides evidence that large non-coding RNAs do not encode proteins. *Cell* **154**: 240–251.

Langfelder P, Horvath S. 2008. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**.

Lee H, Zhang Z, Krause HM. 2019. Long Noncoding RNAs and Repetitive Elements: Junk or Intimate Evolutionary Partners? *Trends Genet* **35**: 892–902.

Li L, Eichten SR, Shimizu R, Petsch K, Yeh C-T, Wu W, Chettoor AM, Givan SA, Cole RA, Fowler JE, et al. 2014. Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biol* **15**.

Li R, Fu D, Zhu B, Luo Y, Zhu H. 2018. CRISPR/Cas9-mediated mutagenesis of lncRNA1459 alters tomato fruit ripening. *Plant J* **94**: 513–524.

Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**: 275–282.

Lin X, Lin W, Ku YS, Wong FL, Li MW, Lam HM, Ngai SM, Chan TF. 2020. Analysis of Soybean Long Non-Coding RNAs Reveals a Subset of Small Peptide-Coding Transcripts. *Plant Physiol* **182**: 1359–1374.

Liu C, Muchhal US, Raghothama KG. 1997. Differential expression of TPS11, a phosphate starvation-induced gene in tomato. *Plant Mol Biol* **33**: 867–874.

Liu F, Marquardt S, Lister C, Swiezewski S, Dean C, Shaw RG, Byers DL, Darmo E, Initiative AG, Lynch M, et al. 2010. Targeted 3′ Processing of Antisense Transcripts Triggers Arabidopsis FLC Chromatin Silencing. *Science (80- )* **327**: 94–98.

Liu F, Xu Y, Chang K, Li S, Liu Z, Qi S, Jia J, Zhang M, Crawford NM, Wang Y, et al. 2019. The long noncoding RNA T5120 regulates nitrate response and assimilation in Arabidopsis. *New Phytol* **1**: 117–131.

Liu HJ, Jian L, Xu J, Zhang Q, Zhang MM, Jin M, Peng Y, Yan JJ, Han B, Liu J, et al. 2020. High-Throughput CRISPR/Cas9 Mutagenesis Streamlines Trait Gene Identification in Maize. *Plant Cell* **32**: 1397–1413.

Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua NH. 2012. Genome-Wide Analysis Uncovers Regulation of Long Intergenic Noncoding RNAs in Arabidopsis.

*Plant Cell* **24**: 4333–4345.

Liu J, Wang H, Chua NH. 2015. Long noncoding RNA transcriptome of plants. *Plant Biotechnol J* **13**: 319–328.

Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, Attenello FJ, Villalta JE, Cho MY, Chen Y, et al. 2017. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science (80- )* **39**.

Liu X, Li D, Zhang D, Yin D, Zhao Y, Ji C, Zhao X, Li X, He Q, Chen R, et al. 2018. A novel antisense long noncoding RNA , TWISTED LEAF , maintains leaf blade flattening by regulating its associated sense R2R3-MYB gene in rice. *New Phytol* **218**: 774–788.

Livak KJ, Schmittgen TD. 2001. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2(-Delta Delta C(T)) Method. **408**: 402–408.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Lu Y, Ye X, Guo R, Huang J, Wang W, Tang J, Tan L, Zhu J kang, Chu C, Qian Y. 2017. Genome-wide Targeted Mutagenesis in Rice Using the CRISPR/Cas9 System. *Mol Plant* **10**: 1242–1245.

Lubelsky Y, Ulitsky I. 2018. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* **555**: 107–111.

Lucero L, Fonouni-Farde C, Crespi M, Ariel F. 2020. Long noncoding RNAs shape transcription in plants. *Transcription* 1–12.

Luo S, Lu JY, Liu L, Yin Y, Chen C, Han X, Wu B, Xu R, Liu W, Yan P, et al. 2016. Divergent lncRNAs regulate gene expression and lineage differentiation in pluripotent cells. *Cell Stem Cell* **18**: 637–652.

Lv Y, Hu F, Zhou Y, Wu F, Gaut BS. 2019. Maize transposable elements contribute to long non-coding RNAs that are regulatory hubs for abiotic stress response. *BMC Genomics* **20**.

Ma J, Yan B, Qu Y, Qin F, Yang Y, Hao X, Yu J, Zhao Q, Zhu D, Ao G. 2008. Zm401, a short open reading frame mRNA or noncoding RNA, is essential for tapetum and microspore development and can regulate the floret formation in maize. *J Cell Biochem* **146**: 136–146.

Ma X, Shao C, Jin Y, Wang H, Meng Y. 2014. Long non-coding RNA: A novel endogenous source for the generation of Dicer-like 1-dependent small RNAs in Arabidopsis thaliana. *RNA Biol* **11**: 373–390.

Macintosh GC, Wilkerson C, Green PJ. 2001. Identification and Analysis of Arabidopsis Expressed Sequence Tags Characteristic of Non-Coding RNAs 1. *Plant Physiol* **127**: 765–776.

Marchese FP, Raimondi I, Huarte M. 2017. The multidimensional mechanisms of long noncoding RNA function. *Genome Biol* **18**: 1–13.

Marquardt S, Raitskin O, Wu Z, Liu F, Sun Q, Dean C. 2014. Functional Consequences of Splicing of the Antisense Transcript COOLAIR on FLC Transcription. *Mol Cell* **54**: 156–165.

Matsui A, Ishida J, Morosawa T, Mochizuki Y, Kaminuma E, Endo TA, Okamoto M, Nambara E, Nakajima M, Kawashima M, et al. 2008. Arabidopsis Transcriptome Analysis under Drought , Cold , High-Salinity and ABA Treatment Conditions using a Tiling Array. *Plant Cell Physiol* **49**: 1135–1149.

Matsui A, Nguyen AH, Nakaminami K, Seki M. 2013. Arabidopsis non-coding RNA regulation in abiotic stress responses. *Int J Mol Sci* **14**: 22642–22654.

Melé M, Mattioli K, Mallard W, Shechner DM, Gerhardinger C, Rinn JL. 2017. Chromatin environment , transcriptional regulation , and splicing distinguish lincRNAs and mRNAs. *Genome Res* **27**: 27–37.

Melo CA, Drost J, Wijchers PJ, van de Werken H, de Wit E, Vrielink JAFO, Elkon R, Melo SA, Léveillé N, Kalluri R, et al. 2013. eRNAs Are Required for p53-Dependent Enhancer Activity and Gene Transcription. *Mol Cell* **49**: 524–535.

Meng X, Yu H, Zhang Y, Zhuang F, Song X, Gao S, Gao C, Li J. 2017. Construction of a Genome-Wide Mutant Library in Rice Using CRISPR/Cas9. *Mol Plant* **10**: 1238–1241.

Michal Z, Fajkus P, Pe V, Fojtov M, Fulne J, Kilar A, Dvoˇ M, Sims J, Eva S. 2019. Telomerase RNAs in land plants. *Nucleic Acids Res* **47**: 9842–9856.

Mohammadin S, Edger PP, Pires JC, Schranz ME. 2015. Positionally-conserved but sequence-diverged: Identification of long non-coding RNAs in the Brassicaceae and Cleomaceae. *BMC Plant Biol* **15**: 217.

Mousavi K, Zare H, Dell'Orso S, Grontved L, Gutierrez-Cruz G, Derfoul A, Hager GL, Sartorelli V. 2013. eRNAs Promote Transcription by Establishing Chromatin Accessibility at Defined Genomic Loci. *Mol Cell* **51**: 606–617.

Mukherjee N, Calviello L, Hirsekorn A, De Pretis S, Pelizzola M, Ohler U, Pretis S De, Pelizzola M, Ohler U. 2017. Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nat Struct Mol Biol* **24**: 86–96.

Nair L, Chung H, Basu U. 2020. Regulation of long non-coding RNAs and genome dynamics by the RNA surveillance machinery. *Nat Rev Mol Cell Biol* **21**: 123–136.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H, Baker JC, et al. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**: 635–640.

Nordin M, Bergman D, Halje M, Engström W, Ward A. 2014. Epigenetic regulation of the Igf2/H19 gene cluster. *Cell Prolif* **47**: 189–199.

Okamoto M, Tatematsu K, Matsui A, Morosawa T, Ishida J, Tanaka M, Endo TA, Mochizuki Y, Toyoda T, Kamiya Y, et al. 2010. Genome-wide analysis of endogenous abscisic acid-mediated transcription in dry and imbibed seeds of Arabidopsis using tiling arrays. *Plant J* **62**: 39–51.

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* **20**: 275.

Ounzain S, Pedrazzini T. 2015. The promise of enhancer-associated long noncoding RNAs in cardiac regeneration. *Trends Cardiovasc Med* **25**: 592–602.

Pachnis V, Belayew A, Tilghman SM. 1984. Locus unlinked to α-fetoprotein under the control of the murine raf and Rif genes. *Proc Natl Acad Sci U S A* **81**: 5523–5527.

Pajoro A, Madrigal P, Muiño JM, Matus JT, Jin J, Mecchia MA, Debernardi JM, Palatnik JF, Balazadeh S, Arif M, et al. 2014. Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. *Genome Biol* **15**: R41.

Paralkar VR, Taborda CC, Huang P, Yao Y, Kossenkov A V., Prasad R, Luan J, Davies JOJ, Hughes JR, Hardison RC, et al. 2016. Unlinking an lncRNA from Its Associated cis Element. *Mol Cell* **62**: 104–110.

Pefanis E, Wang J, Rothschild G, Lim J, Kazadi D, Sun J, Federation A, Chao J, Elliott O, Liu ZP, et al. 2015. RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. *Cell* **161**: 774–789.

Pegueroles C, Iraola-Guzmán S, Chorostecki U, Ksiezopolska E, Saus E, Gabaldón T. 2019. Transcriptomic analyses reveal groups of co-expressed, syntenic lncRNAs in four species of the genus Caenorhabditis. *RNA Biol* **16**: 320–329.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290–295.

Plewka P, Thompson A, Szymanski M, Nuc P, Knop K, Rasinska A, Bialkowska A, Szweykowska-Kulinska Z, Karlowski WM, Jarmolowski A. 2018. A stable tRNA-like molecule is generated from the long noncoding RNA GUT15 in Arabidopsis. *RNA Biol* **6286**: 726–738.

Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. 2019. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol* **20**:

38.

Qin T, Zhao H, Cui P, Albesher N, Xionga L, Xiong L. 2017. A Nucleus-Localized Long Non-Coding RNA Enhances Drought and Salt Stress Tolerance. *Plant Physiol* **175**: 1321–1336.

Quinn JJ, Chang HY. 2016. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet* **17**: 47–62.

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**: W160–W165.

Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, et al. 2007. Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell* **129**: 1311–1323.

Rosa S, Duncan S, Dean C. 2016. Mutually exclusive sense–antisense transcription at FLC facilitates environmentally induced gene repression. *Nat Commun* **7**: 1–7.

Ruiz-Orera J, Mar M, Albà AA. 2019. Conserved regions in long non-coding RNAs contain abundant translation and protein-RNA interaction signatures. *NAR Genomics Bioinforma* **1**: 2.

Sarropoulos I, Marin R, Cardoso-Moreira M, Kaessmann H. 2019. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571**: 510–514.

Sela N, Kim E, Ast G. 2010. The role of transposable elements in the evolution of non-mammalian vertebrates and invertebrates. *Genome Biol* **11**: 59.

Selleri L, Bartolomei MS, Bickmore WA, He L, Stubbs L, Reik W, Barsh GS. 2016. A Hox-Embedded Long Noncoding RNA: Is It All Hot Air?. *PLOS Genet* **12**: 8–12.

Seo JS, Chua NH, Fluorescence T, Trifc C. 2019. Trimolecular Fluorescence Complementation (TriFC) Assay for Visualization of RNA-Protein Interaction in Plants. *Methods Mol Biol* **1933**: 297–303.

Seo JS, Sun HX, Park BS, Huang CH, Yeh SD, Jung C, Chua NH. 2017. ELF18-INDUCED LONG-NONCODING RNA Associates with Mediator to Enhance Expression of Innate Immune Response Genes in Arabidopsis. *Plant Cell* **29**: 1024–1038.

Sequeira-Mendes J, Araguez I, Peiro R, Mendez-Giraldez R, Zhang X, Jacobsen SE, Bastolla U, Gutierrez C. 2014. The Functional Topography of the Arabidopsis Genome Is Organized in a Reduced Number of Linear Motifs of Chromatin States. *Plant Cell* **26**: 2351–2366.

Severing E, Faino L, Jamge S, Busscher M, Kuijer-Zhang Y, Bellinazzo F, Busscher-Lange J, Fernández V, Angenent GC, Immink RGH, et al. 2018. Arabidopsis thaliana ambient temperature responsive lncRNAs. *BMC Plant Biol* **18**: 145.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.

Shepard PJ, Choi EA, Lu J, Flanagan LA, Hertel KJ, Shi Y. 2011. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA* **17**: 761–772.

Shin JH, Chekanova JA. 2014. Arabidopsis RRP6L1 and RRP6L2 Function in FLOWERING LOCUS C Silencing via Regulation of Antisense RNA Synthesis. *PLoS Genet* **10**: 21–24.

Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: From properties to genome-wide predictions. *Nat Rev Genet* **15**: 272–286.

Smaczniak C, Immink RGH, Muiño JM, Blanvillain R, Busscher M, Busscher-Lange J, Dinh QD, Liu S, Westphal AH, Boeren S, et al. 2012. Characterization of MADS-domain transcription factor complexes in Arabidopsis flower development. *Proc Natl Acad Sci U S A* **109**: 1560–1565.

Song J, Logeswaran D, Castillo-González C, Li Y, Bose S, Aklilu BB, Ma Z, Polkhovskiy A, Chen JJL, Shippen DE. 2019. The conserved structure of plant telomerase RNA provides the missing link for an evolutionary pathway from ciliates to humans. *Proc Natl Acad Sci U S A* **116**: 24542–24550.

Song JH, Cao JS, Wang CG. 2013. BcMF11, a novel non-coding RNA gene from Brassica campestris, is required for pollen development and male fertility. *Plant Cell Rep* **32**: 21–30.

Spector DL, Lamond AI. 2011. Nuclear speckles. *Cold Spring Harb Perspect Biol* **3**: 1–12.

Swiezewski S, Liu F, Magusin A, Dean C. 2009. Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature* **462**: 799–802.

Szabo EX, Reichert P, Lehniger M-K, Ohmer M, de Francisco Amorim M, Gowik U, Schmitz-Linneweber C, Laubinger S. 2020. Metabolic Labeling of RNAs Uncovers Hidden Features and Dynamics of the Arabidopsis Transcriptome. *Plant Cell* **32**: tpc.00214.2019.

Szcześniak MW, Rosikiewicz W, Makałowska I. 2016. CANTATAdb: A collection of plant long non-coding RNAs. *Plant Cell Physiol* **57**: e8.

The UniProt Consortium. 2017. UniProt : the universal protein knowledgebase. *Nucleic Acids Res* **45**: 158–169.

Thieffry A, Vigh ML, Bornholdt J, Ivanov M, Brodersen P, Sandelin A. 2020a. Characterization of Arabidopsis thaliana promoter Bidirectionality and Antisense RNAs by Depletion of Nuclear RNA Decay Pathways. *Plant Cell* **32**: 1845–1867.

Thieffry A, Vigh ML, Bornholdt J, Ivanov M, Brodersen P, Sandelin A. 2020b. Characterization of Arabidopsis thaliana Promoter Bidirectionality and Antisense RNAs by Inactivation of Nuclear RNA Decay Pathways. *Plant Cell* **32**: 1845–1867.

Tian C, Wang Y, Yu H, He J, Wang J, Shi B, Du Q, Provart NJ, Meyerowitz EM, Jiao Y. 2019. A gene expression map of shoot domains reveals regulatory mechanisms. *Nat Commun* **10**: 1–12.

Ulitsky I. 2016. Evolution to the rescue: Using comparative genomics to understand long non-coding RNAs. *Nat Rev Genet* **17**: 601–614.

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**: 1537–1550.

Van de Velde J, Van Bel M, Vaneechoutte D, Vandepoele K. 2016. A collection of conserved noncoding sequences to study gene regulation in flowering plants. *Plant Physiol* **171**: 2586–2598.

Wang D, Qu Z, Yang L, Zhang Q, Liu ZH, Do T, Adelson DL, Wang ZY, Searle I, Zhu JK. 2017a. Transposable elements (TEs) contribute to stress-related long intergenic noncoding RNAs in plants. *Plant J* **90**: 133–146.

Wang H, Chung PJ, Liu J, Jang IC, Kean MJ, Xu J, Chua NH. 2014a. Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. *Genome Res* **24**: 444–453.

Wang H, Niu QW, Wu HW, Liu J, Ye J, Yu N, Chua NH. 2015a. Analysis of non-coding transcriptome in rice and maize uncovers roles of conserved lncRNAs associated with agriculture traits. *Plant J* **84**.

Wang J, Meng X, Dobrovolskaya OB, Orlov YL, Chen M. 2017b. Non-coding RNAs and Their Roles in Stress Response in Plants. *Genomics, Proteomics Bioinforma* **15**: 301–312.

Wang J, Yu W, Yang Y, Li X, Chen T, Liu T, Ma N, Yang X, Liu R, Zhang B. 2015b. Genome-wide analysis of tomato long non-coding RNAs and identification as endogenous target mimic for microRNA in response to TYLCV infection. *Sci Rep* **5**: 1–16.

Wang KC, Chang HY. 2011. Molecular Mechanisms of Long Noncoding RNAs. *Mol Cell* **43**: 904–914.

Wang M, Yuan D, Tu L, Gao W, He Y, Hu H, Wang P, Liu N, Lindsey K, Zhang X. 2015c. Long noncoding RNAs and their proposed functions in fibre development of cotton (Gossypium spp.). *New Phytol* **207**: 1181–1197.

Wang HL V., Chekanova JA. 2019. Novel mRNAs 3′ end-associated cis-regulatory elements with epigenomic signatures of mammalian enhancers in the Arabidopsis genome. *RNA* **25**:

1242–1258.

Wang X, Ai G, Zhang C, Cui L, Wang J, Li H, Zhang J, Ye Z. 2015d. Expression and diversification analysis reveals transposable elements play important roles in the origin of Lycopersicon- specific lncRNAs in tomato. *New Phytol* **209**: 1442–1455.

Wang Y, Fan X, Lin F, He G, Terzaghi W, Zhu D, Deng XW. 2014b. Arabidopsis noncoding RNA mediates control of photomorphogenesis by red light. *Proc Natl Acad Sci* **111**: 10359–10364.

Wang Y, Luo X, Sun F, Hu J, Zha X, Su W, Yang J. 2018. Overexpressing lncRNA LAIR increases grain yield and regulates neighbouring gene cluster expression in rice. *Nat Commun* **9**: 1–9.

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**: 1–14.

Wasaki J, Yonetani R, Shinano T, Kai M, Osaki M. 2003. Expression of the OsPI1 gene , cloned from rice roots using cDNA microarray , rapidly responds to phosphorus status. 239–248.

Washietl S, Findeiß S, Müller SA, Kalkhof S, Von Bergen M, Hofacker IL, Stadler PF, Goldman N. 2011. RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* **17**: 578–594.

Washietl S, Kellis M, Garber M. 2014. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res* **24**: 616–628.

Wei S, Chen H, Dzakah EE, Yu B, Wang X, Fu T, Li J, Liu L, Fang S, Liu W, et al. 2019. Systematic evaluation of C . elegans lincRNAs with CRISPR knockout mutants. *Genome Biol* **20**: 1–19.

Wibowo A, Becker C, Marconi G, Durr J, Price J, Hagmann J, Papareddy R, Putra H, Kageyama J, Becker J, et al. 2016. Hyperosmotic stress memory in Arabidopsis is mediated by distinct epigenetically labile sites in the genome and is restricted in the male germline by DNA glycosylase activity. *Elife* **5**: 1–27.

Wilusz JE, Freier SM, Spector DL. 2008. 3' end processing of a long nuclear-retained non-coding RNA yields a tRNA-like cytoplasmic RNA. *Cell* **135**: 919–932.

Wu HW, Deng S, Xu H, Mao HZ, Liu J, Niu QW, Wang H, Chua NH. 2018. A noncoding RNA transcribed from the AGAMOUS (AG) second intron binds to CURLY LEAF and represses AG expression in leaves. *New Phytol* **219**: 1480–1491.

Wu J, Liu C, Liu Z, Li S, Li D, Liu SS, Huang X, Liu SS, Yukawa Y. 2019. Pol III-Dependent Cabbage BoNR8 Long ncRNA Affects Seed Germination and Growth in Arabidopsis. *Plant Cell Physiol* **60**: 421–435.

Wu J, Okada T, Fukushima T, Tsudzuki T, Sugiura M, Yukawa Y. 2012. A novel hypoxic stress-responsive long non-coding RNA transcribed by RNA polymerase III in Arabidopsis.Telomerase RNAs in land plants. *RNA Biol* **6286**: 302–313.

Wu Z, Fang X, Zhu D, Dean C. 2020. Autonomous Pathway: FLOWERING LOCUS C Repression through an Antisense-Mediated Chromatin-Silencing Mechanism. *Plant Physiol* **182**: 27–37.

Wunderlich M, Groß-Hardt R, Schöffl F. 2014. Heat shock factor HSFB2a involved in gametophyte development of Arabidopsis thaliana and its expression is controlled by a heat-inducible long non-coding antisense RNA. *Plant Mol Biol* **85**: 541–550.

Xie Y, Liu Y, Ma M, Zhou Q, Zhao Y, Zhao B, Wang B, Wei H, Wang H. 2020. Arabidopsis FHY3 and FAR1 integrate light and strigolactone signaling to regulate branching. *Nat Commun* **11**: 1–13.

Yan W, Chen D, Kaufmann K. 2016. Molecular mechanisms of floral organ specification by MADS domain proteins. *Curr Opin Plant Biol* **29**: 154–162.

Yan W, Chen D, Schumacher J, Durantini D, Engelhorn J, Chen M, Carles CC, Kaufmann K. 2019. Dynamic control of enhancer activity drives stage-specific gene expression during flower

morphogenesis. *Nat Commun* **10**: 1–16.

Yang T, Ma H, Zhang J, Wu T, Song T, Tian J, Yao Y. 2019a. Systematic identification of long noncoding RNAs expressed during light-induced anthocyanin accumulation in apple fruit. *Plant J* **100**: 572–590.

Yang XH, Nadadur RD, Hilvering CR, Bianchi V, Werner M, Mazurek SR, Gadek M, Shen KM, Goldman JA, Tyan L, et al. 2017. Transcription-factor-dependent enhancer transcription defines a gene regulatory network for cardiac rhythm. *Elife* **6**.

Yang Y, Liu T, Shen D, Wang J, Ling X, Hu Z, Chen T, Hu J, Huang J, Yu W, et al. 2019b. Tomato yellow leaf curl virus intergenic siRNAs target a host long noncoding RNA to modulate disease symptoms ed. S.P. Dinesh-Kumar. *PLOS Pathog* **15**: 1–22.

Yin Q, Yang L, Zhang Y, Xiang J, Wu Y, Carmichael GG, Chen L. 2012. Long Noncoding RNAs with snoRNA Ends. *Mol Cell* **48**: 219–230.

Yin Y, Lu JY, Zhang X, Shao W, Xu Y, Li P, Hong Y, Cui L, Shan G, Tian B, et al. 2020. U1 snRNP regulates chromatin retention of noncoding RNAs. *Nature* **580**: 147–150.

Yin Y, Yan P, Lu J, Song G, Zhu Y, Li Z, Zhao Y, Shen B, Huang X, Zhu H, et al. 2015. Opposing roles for the lncRNA haunt and its genomic locus in regulating HOXA gene activation during embryonic stem cell differentiation. *Cell Stem Cell* **16**: 504–516.

You Y, Sawikowska A, Neumann M, Posé D, Capovilla G, Langenecker T, Neher RA, Krajewski P, Schmid M. 2017. Temporal dynamics of gene expression and histone marks at the Arabidopsis shoot meristem during flowering. *Nat Commun* **8**: 1–12.

Young MD, Wakefield MJ, Smyth GK, Oshlack A. 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11**: R14.

Yu Y, Zhang Y, Chen X, Chen Y. 2019a. Plant Noncoding RNAs: Hidden Players in Development and Stress Responses. *Annu Rev Cell Dev Biol* **35**: 407–431.

Yu Y, Zhou YF, Feng YZ, He H, Lian JP, Yang YW, Lei MQ, Zhang YC, Chen YQ, He H, et al. 2019b. Transcriptional landscape of pathogen-responsive lncRNAs in rice unveils the role of ALEX1 in jasmonate pathway and disease resistance. *Plant Biotechnol J* **18**: 1–12.

Yuan J, Li J, Yang Y, Tan C, Zhu Y, Hu L, Qi Y, Lu ZJ. 2018. Stress-responsive regulation of long non-coding RNA polyadenylation in Oryza sativa. *Plant J* **93**: 814–827.

Yuan J, Zhang Y, Dong J, Sun Y, Lim BL, Liu D, Lu ZJ. 2016. Systematic characterization of novel lncRNAs responding to phosphate starvation in Arabidopsis thaliana. *BMC Genomics* **17**: 655.

Zhang B, Pan X, Cannon CH, Cobb GP, Anderson TA. 2006. Conservation and divergence of plant microRNA genes. *Plant J* **46**: 243–259.

Zhang G, Chen D, Zhang T, Duan A, Zhang J, He C. 2018a. Transcriptomic and functional analyses unveil the role of long non-coding RNAs in anthocyanin biosynthesis during sea buckthorn fruit ripening. **25**: 465–476.

Zhang H, Tao Z, Hong H, Chen Z, Wu C, Li X, Xiao J, Wang S. 2016. Transposon-derived small RNA is responsible for modified function of WRKY45 locus. *Nat Plants* **2**: 1–8.

Zhang K, Wang X, Cheng F. 2019a. Plant Polyploidy: Origin, Evolution, and Its Influence on Crop Domestication. *Hortic Plant J* **5**: 231–239.

Zhang L, Wang M, Li N, Wang H, Qiu P, Pei L, Xu Z, Wang T, Gao E, Liu J, et al. 2018b. Long noncoding RNAs involve in resistance to Verticillium dahliae , a fungal disease in cotton. *Plant Biotechnol J* **16**: 1172–1185.

Zhang P, Du H, Wang J, Pu Y, Yang C, Yan R, Yang H, Cheng H, Yu D. 2020. Multiplex CRISPR/Cas9-mediated metabolic engineering increases soya bean isoflavone content and resistance to soya bean mosaic virus. *Plant Biotechnol J* **18**: 1384–1395.

Zhang X, Dong J, Deng F, Wang W, Cheng Y, Song L, Hu M, Shen J, Xu Q, Shen F. 2019b. The long non-coding RNA lncRNA973 is involved in cotton response to salt stress. *BMC Plant Biol* **19**: 1–16.

Zhang Y-C, Liao J-Y, Li Z-Y, Yu Y, Zhang J-P, Li Q-F, Qu L-H, Shu W-S, Chen Y. 2014. Genome-

wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. *Genome Biol* **15**: 512.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zhao T, Tao X, Feng S, Wang L, Hong H, Ma W, Shang G, Guo S. 2018a. LncRNAs in polyploid cotton interspecific hybrids are derived from transposon neofunctionalization. 1–17.

Zhao X, Li J, Lian B, Gu H, Li Y, Qi Y. 2018b. Global identification of Arabidopsis lncRNAs reveals the regulation of MAF4 by a natural antisense RNA. *Nat Commun* **9**: 1–12.

Zhou B, Zhao H, Yu J, Guo C, Dou X, Song F, Hu G, Cao Z, Qu Y, Yang Y, et al. 2018. EVLncRNAs: a manually curated database for long non-coding RNAs validated by low-throughput. *Nucleic Acids Res* **46**: 100–105.

Zhou B, Zhao H, Yu J, Guo C, Dou X, Song F, Hu G, Cao Z, Qu Y, Yang Y, et al. 2019. Experimentally Validated Plant lncRNAs in EVLncRNAs Database. *Methods Mol Biol* **1933**: 431–437.

Zhou H, Liu Q, Li J, Jiang D, Zhou L, Wu P, Lu S, Li F, Zhu L, Liu Z, et al. 2012. Photoperiod- and thermo-sensitive genic male sterility in rice are caused by a point mutation in a novel noncoding RNA that produces a small RNA. *Cell Res* **22**: 649–660.

Zhu B, Yang Y, Li R, Fu D, Wen L, Luo Y, Zhu H. 2015a. RNA sequencing and functional analysis implicate the regulatory role of long non-coding RNAs in tomato fruit ripening. *J Exp Bot* **66**: 4483–4495.

Zhu B, Zhang W, Zhang T, Liu B, Jiang J. 2015b. Genome-wide prediction and validation of intergenic enhancers in arabidopsis using open chromatin signatures. *Plant Cell* **27**: 2415–2426.

Zhu QH, Stephen S, Taylor J, Helliwell CA, Wang MB. 2014. Long noncoding RNAs responsive to Fusarium oxysporum infection in Arabidopsis thaliana. *New Phytol* **201**: 574–584.

Zhu Y, Rowley MJ, Böhmdorfer G, Wierzbicki AT. 2013. A SWI/SNF Chromatin-Remodeling Complex Acts in Noncoding RNA-Mediated Transcriptional Silencing. *Mol Cell* **49**: 298–309.

**Supplemental data**

A

| study | Liu et al, 2012 | published datasets |
|---|---|---|
| total number | 6480 | 19740 |
| overlap number | 332 | 1311 |
| overlapping rates | 10% | 38% |

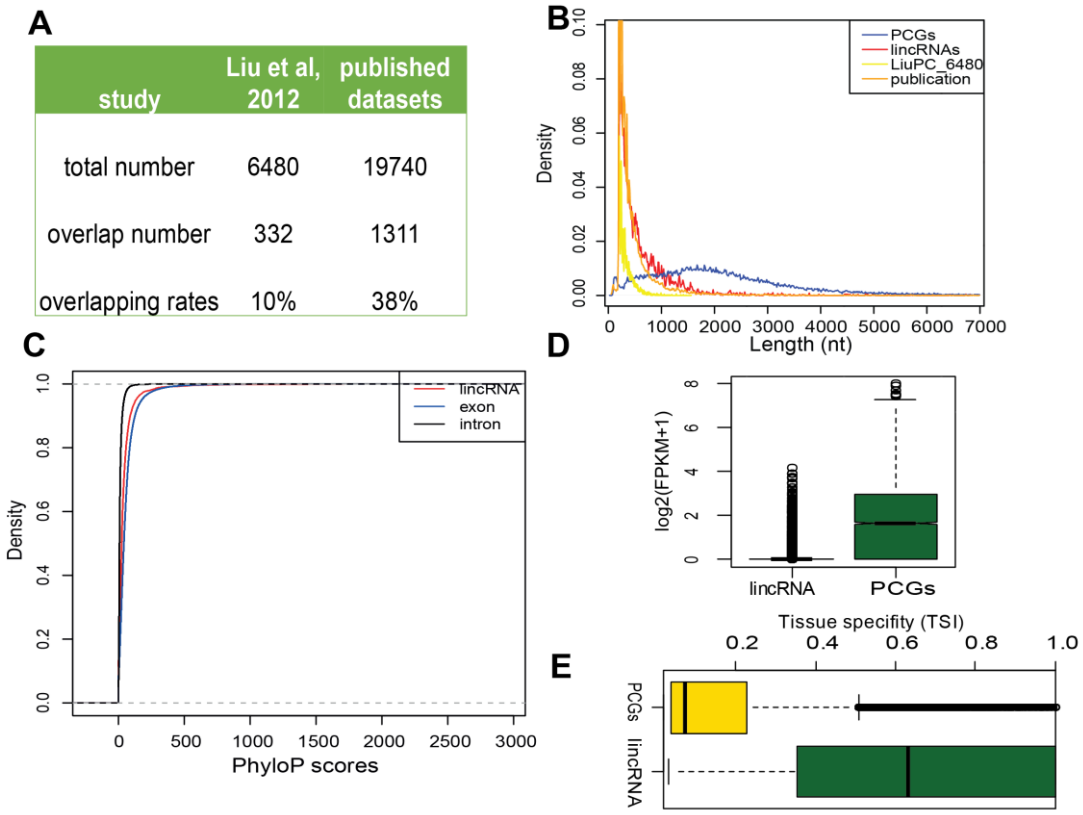



**Figure S1**: **The characteristics of our identified lincRNAs**. (A) Comparison of our identified lincRNAs in this study with 6480 published lincRNAs in Liu et al, 2012, and 19740 published lincRNAs show a large part of our lincRNAs are novel. The comparison was done by the bedtools with at least 1nt overlapping. (B) Length distribution of our identified lincRNAs, published lincRNAs, and PCGs. (C) Evolutionary conservation (PhyloP score) of lincRNA, exons, and introns of PCGs. (D) Distribution of expression levels of lincRNAs and PCGs. (E) The expression specificity (TSI) of lincRNAs and PCGs. The TSI value of 1 represents tissue-specific expression, while the TSI value of 0 represents constant such as housekeeping genes due to constitutive expression in all tissues.

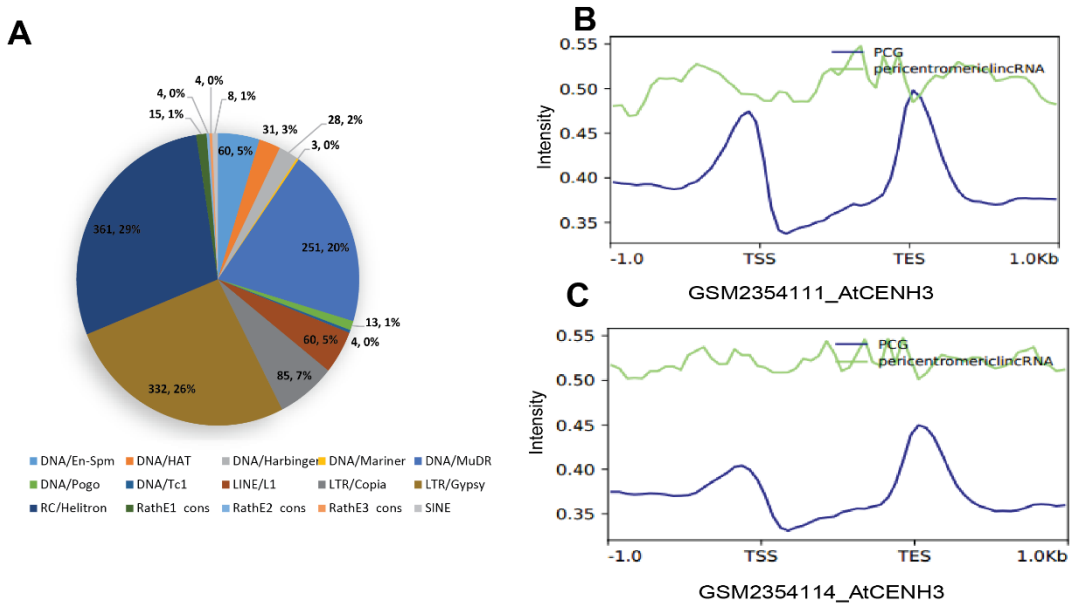
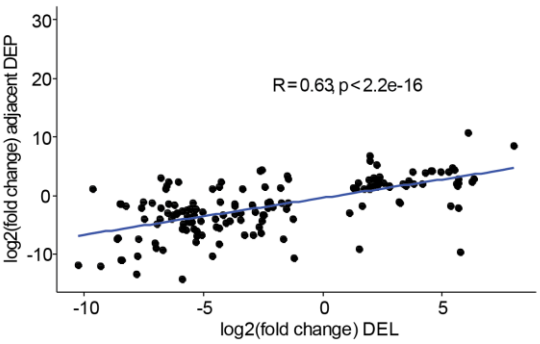



**Figure S2**: **Non-random spatial distribution of lincRNAs in the *Arabidopsis* genome.** (A) Transposable element family/subfamily composition of pericentromeric lincRNAs. (B) Distribution of CENH3 (GSM2354111_AtCENH3) around TSSs and TESs of pericentromeric lincRNAs and PCGs. (C) Distribution of CENH3 (GSM2354114_AtCENH3) around TSSs and TESs of pericentromeric lincRNAs and PCGs.

**A**

| Cmp_groups | lincRNAs | PCGs |
|---|---|---|
| polyA-Stage0_vs_polyA-Stage2 | 5(1%) | 304(1%) |
| polyA-Stage0_vs_polyA-Stage4 | 31(4%) | 948(3%) |
| polyA-Stage0_vs_polyA-Stage8 | 234(29%) | 6187(23%) |
| polyA-Stage2_vs_polyA-Stage4 | 11(1%) | 314(1%) |
| polyA-Stage2_vs_polyA-Stage8 | 168(21%) | 5721(21%) |
| polyA-Stage4_vs_polyA-Stage8 | 188(23%) | 5149(19%) |
| totalRNA-stage0_vs_totalRNA-stage2 | 18(2%) | 978(4%) |
| totalRNA-stage0_vs_totalRNA-stage4 | 39(5%) | 1385(5%) |
| totalRNA-stage0_vs_totalRNA-stage8 | 203(25%) | 5387(20%) |
| totalRNA-stage2_vs_totalRNA-stage4 | 2(0%) | 243(1%) |
| totalRNA-stage2_vs_totalRNA-stage8 | 196(24%) | 5293(19%) |
| totalRNA-stage4_vs_totalRNA-stage8 | 183(23%) | 5064(18%) |

**B**




**C**

| | | PCG | |
|---|---|---|---|
| t2_H3K27me3 | | 1 | 0 |
| lincRNA | 1 | 1159 | 0 |
| | 0 | 0 | 2334 |

p-value=0

| | | PCG | |
|---|---|---|---|
| t2_H3K4me3 | | 1 | 0 |
| lincRNA | 1 | 139 | 0 |
| | 0 | 0 | 3354 |

p-value=4.86E-253

**Figure S3**: **Differential expression of lincRNAs.** (A) Differential expression lincRNAs and PCGs between stages. (B) Scatterplot displaying the relationship between the fold change (log2) for differential expression of lincRNAs with fold changes for the nearest neighboring protein coding genes. (C) The nearest neighbor target PCGs of lincRNAs are often repressed (H3K27me3, upper panel) or activated (H3K4me3, lower panel) simultaneously.

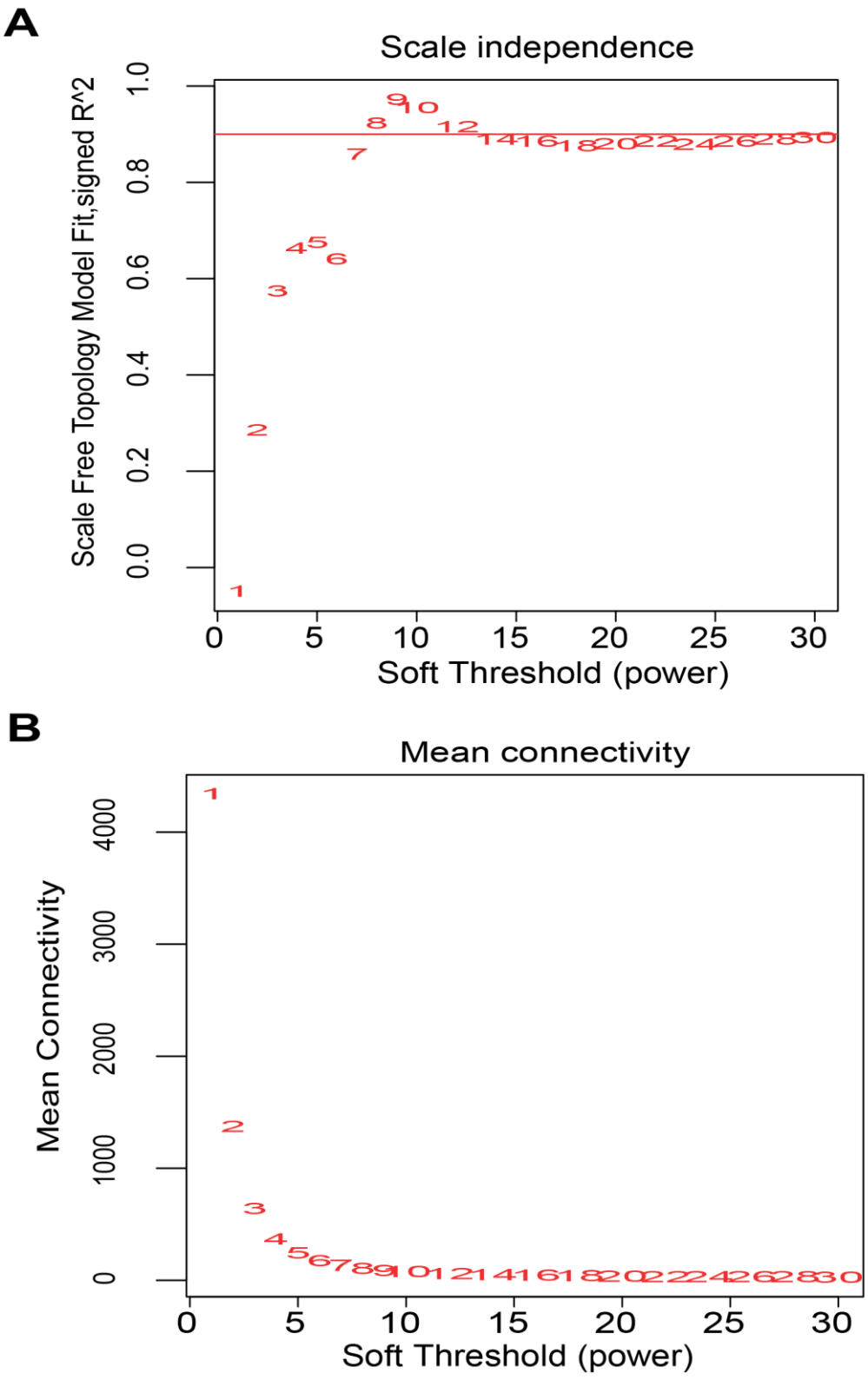



**Figure S4**: **Functional annotation of lincRNAs by lincRNA-PCG co-expression network** (Guilt-by-Association). (A) Scale independence for the lincRNA-PCGs co-expression network by WGCNA. (B) Mean connectivity for the lincRNA-PCGs co-expression network by WGCNA.

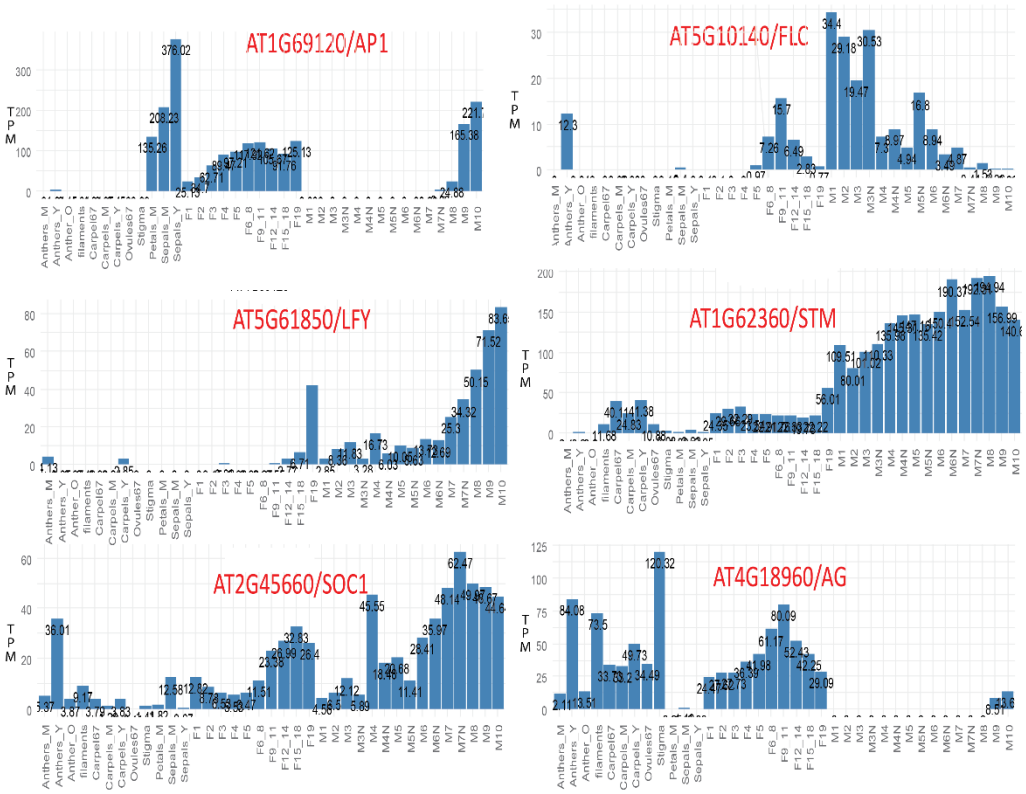



**Figure S5**: Expression pattern of marker genes *AP1*, *FLC*, *LFY, STM, SOC1, and AG* during flowering and floral development.

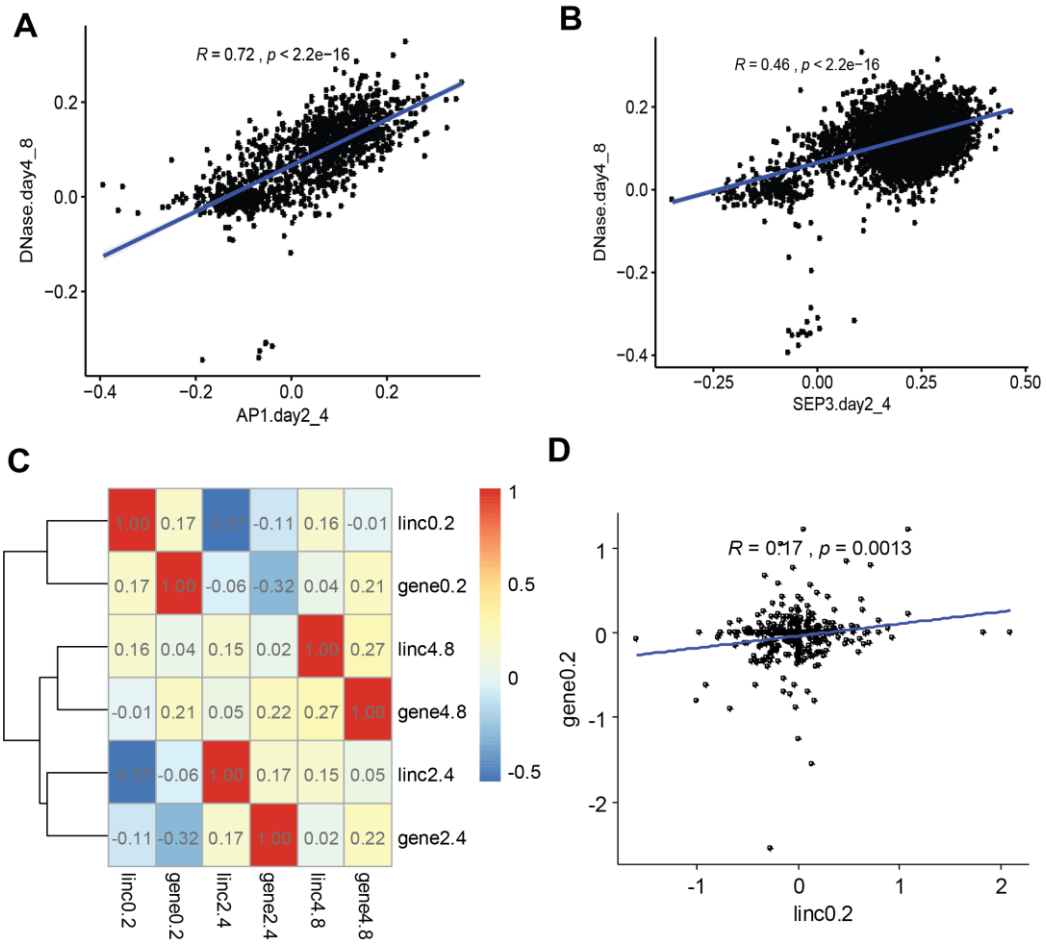
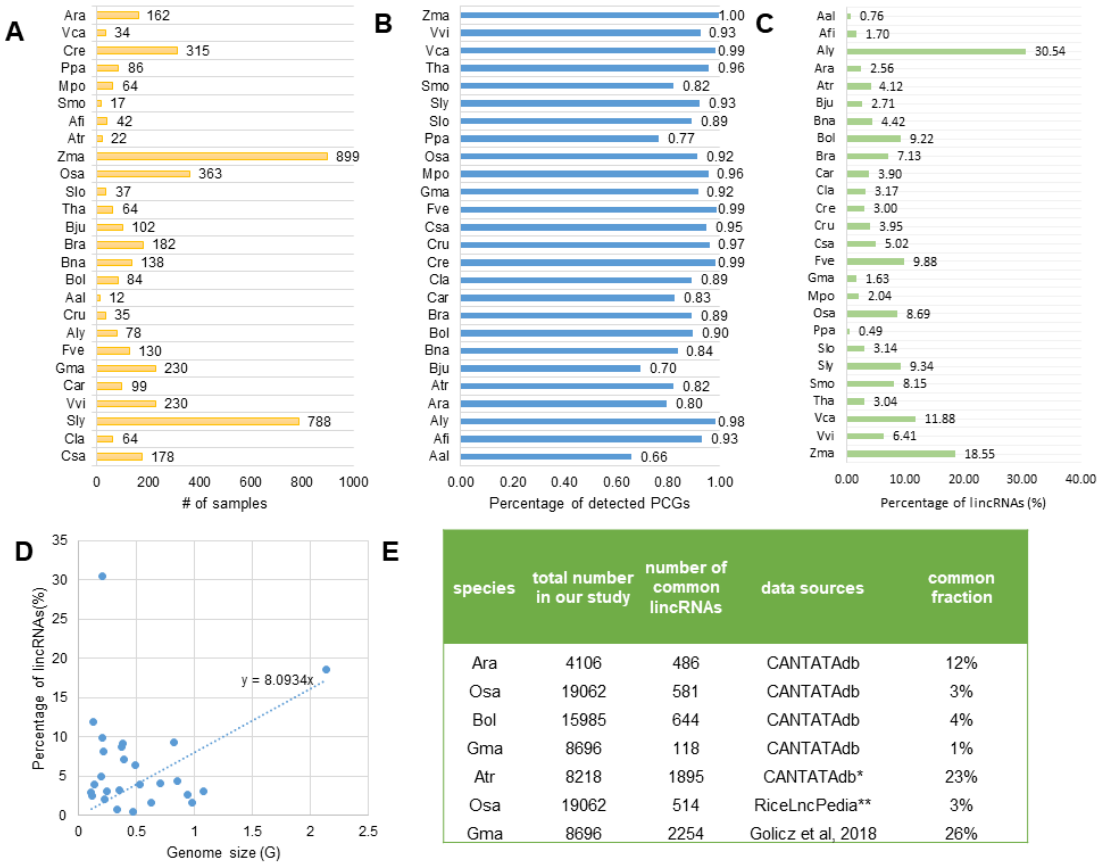
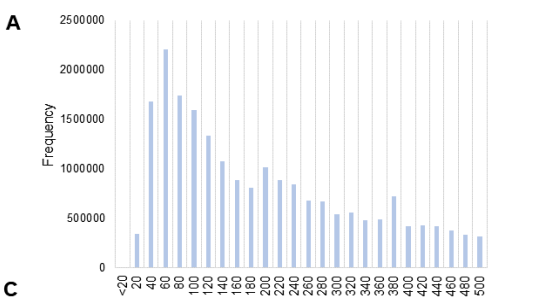



**Figure S6**: **Flower related lincRNAs are the components of floral gene regulatory network. (A)** Regression lines with Pearson correlation coefficients between log2 (FC) change in *AP1* binding from day 2 to day 4 (AP1.day2_4, the same time point) and log2 (FC) change in chromatin accessibility from day 4 to day 8 (DNase.day4_8, the later time point). (**B**) Regression lines with Pearson correlation coefficients between log2 (FC) change in SEP3 binding from day 2 to day 4 (SEP3.day2_4, the same time point) and log2 (FC) change in chromatin accessibility from day 4 to day 8 (DNase.day4_8, the later time point). (**C**) A positive correlation between enhancer associated lincRNAs expression dynamics and that of neighboring protein coding genes**.** Pearson correlation coefficients values between log2(FC) change in enhancer associated lincRNAs expression from the different time points (linc0.2, the previous time point; linc2.4, the same time point; linc4.8, the later time point) and log2(FC) change in expression of the neighboring target genes from the different time points (gene0.2, the previous time point; gene2.4, the same time point; gene4.8, the later time point) are showed in the heatmap. (**D**) Regression lines with Pearson correlation coefficients between log2 (FC) change in expression of enhancers associated lincRNAs from day 0 to day 2 (linc0.2, the previous time point) and log2 (FC) change in the neighboring target genes from day 0 to day 2 (gene0.2, the same time point).




| species | total number in our study | number of common lincRNAs | data sources | common fraction |
|---|---|---|---|---|
| Ara | 4106 | 486 | CANTATAdb | 12% |
| Osa | 19062 | 581 | CANTATAdb | 3% |
| Bol | 15985 | 644 | CANTATAdb | 4% |
| Gma | 8696 | 118 | CANTATAdb | 1% |
| Atr | 8218 | 1895 | CANTATAdb* | 23% |
| Osa | 19062 | 514 | RiceLncPedia** | 3% |
| Gma | 8696 | 2254 | Golicz et al, 2018 | 26% |

**Figure S7**: **Characteristics of identified lincRNAs in each species**. (A) The number of samples used for the identification of lincRNAs in each species. (B) The proportion of expressed PCGs in samples for the identification of lincRNAs in each species. (C) The proportion of lincRNAs covering the genome in each species. (D) The percentage of lincRNAs versus genome size. (E) Overlapping lincRNAs with ones in the publication and public database. *CANTATAdb: http://cantata.amu.edu.pl/; **RiceLncPedia: http://218.199.68.191:10092/.




C

| Species | Num of lincRNAs | overlap | odds ratio | Pvalue |
|---|---|---|---|---|
| *Arabidopsis thaliana* | 4106 | 1017 | 9.65 | 2.09E-200 |
| *Vitis vinifera* | 4896 | 2063 | 6.58 | 1.92E-32 |
| *Brassica rapa* | 13176 | 8095 | 8.61 | 7.43E-257 |
| *Glycine max* | 8700 | 3375 | 23.85 | 0.004946 |
| *Solanum lycopersicum* | 57903 | 5063 | 0.94 | 9.55E-44 |

D

| Species | Num of lincRNAs | overlap | odds ratio | Pvalue |
|---|---|---|---|---|
| *Arabidopsis thaliana* | 4106 | 1717 | 16.29 | 2.25E-29 |
| *Arabidopsis lyrata* | 18012 | 11548 | 2.1 | 2.42E-209 |
| *Capsella rubella* | 2695 | 955 | 8.97 | 8.14E-55 |

E

| Species | Num of lincRNAs | overlap | odds ratio | Pvalue |
|---|---|---|---|---|
| *Arabidopsis thaliana* | 4106 | 718 | 6.81 | 1.82E-297 |

**Figure S8**: **conserved regions within lincRNAs between species**. (A) Frequency distribution of length of alignment length (sequence overlapped) by BLAST 2.4.0+. (B) Frequency distribution of the number of mismatches within one alignment (one blast hit). (C)The overlapping between lincRNAs and conserved non-coding sequences (CNSs) in Van de Velde et al. 2016. (D) The overlapping between lincRNAs and conserved non-coding sequences (CNSs) in Van de Velde et al. 2014. (E) The overlapping between lincRNAs and conserved non-coding sequences (CNSs) in Annabelle Haudry et al, 2013.

| sequence based homolog shared by # of species | species list | family type | Ath | Aal | Aly | Cru | Tha | Bra | Bna | Bol | Bju | Fve | Vvi | Car | Csa | Cla | Gma | Sly | Slo | Atr | Zma | Osa | Ppa | Mpo | Cre | Vca | Smo | Afi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | Afi,Aly,Ath,Atr,Bju,Bna,Bol,Bra,Car,Cla,Csa,Fve,Gma,Osa,Slo,Sly,Tha,Vvi,Zma | many2many | 2 | 0 | 2 | 0 | 3 | 6 | 2 | 5 | 6 | 2 | 3 | 6 | 2 | 6 | 6 | 13 | 4 | 10 | 28 | 1 | 0 | 0 | 0 | 0 | 0 | 4 |
| 3 | Aal,Aly,Ath | one2many | 1 | 1 | 59 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | Bna,Bra,Cla,Cru,Mpo,Slo,Sly,Smo,Vvi,Zma | one2one | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |

**Figure S9**: One typical example of one2one, one2many, many2many lincRNA family, respectively.

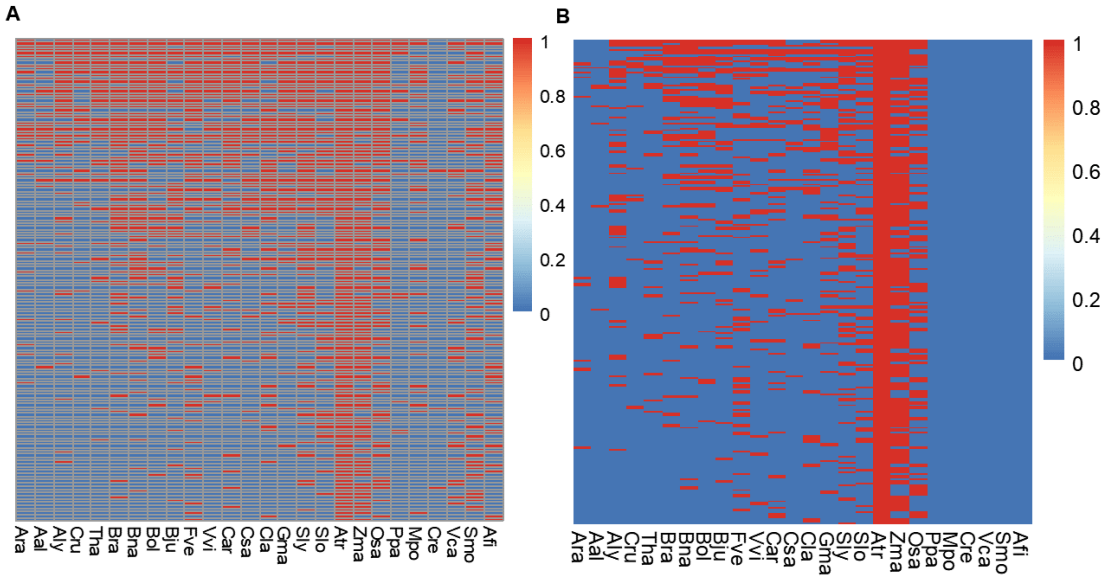



**Figure S10**: **conserved lincRNAs in plants**. (A) The 152 plants (the evolutionary age group: Plants) lincRNA families (27 one2one lincRNA family). (B) The 217 flowering plants (the evolutionary age group: Angiosperms) lincRNA families (94 one2one lincRNA family).

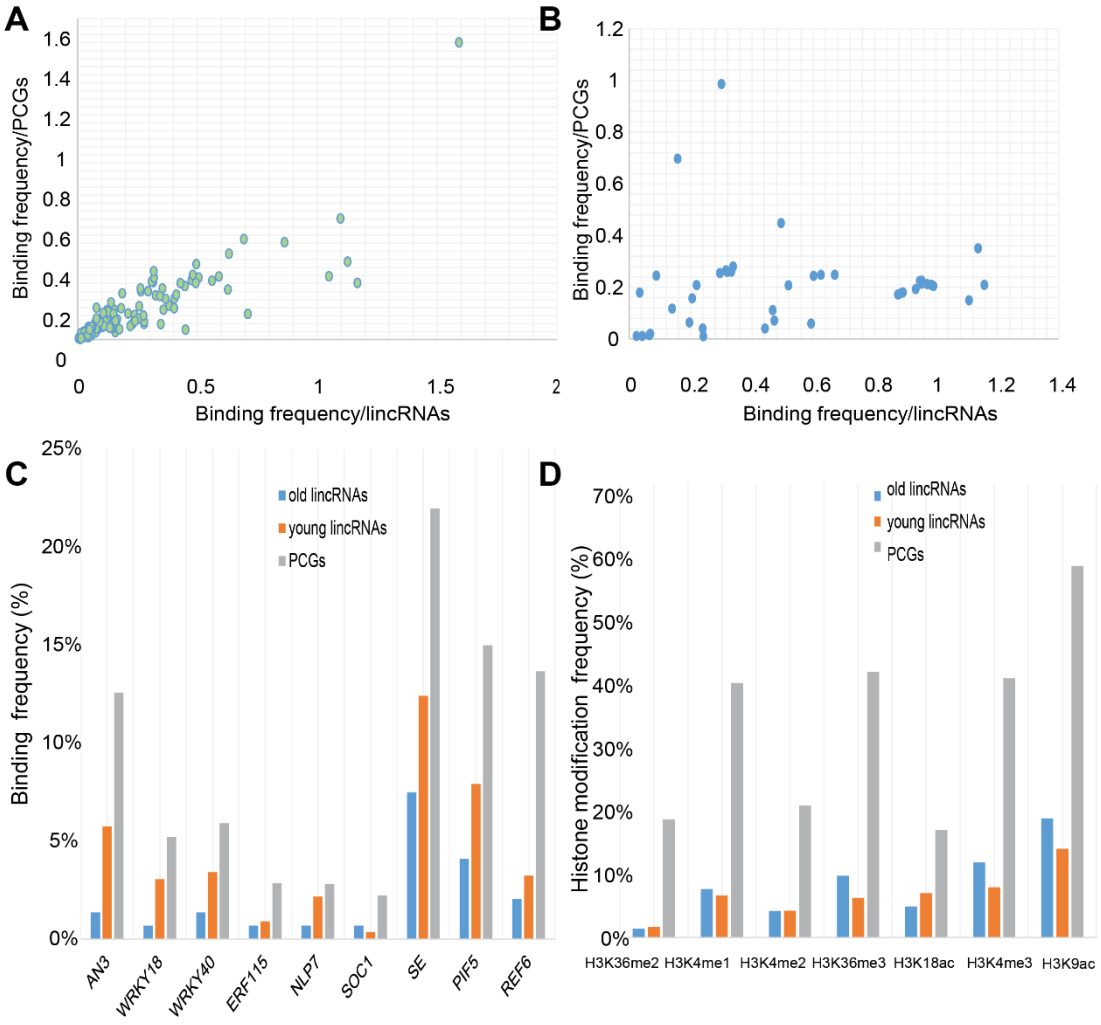



**Figure S11**: **evidence for regulation of conserved lincRNAs in plants**. (A) Comparison between the frequencies of transcription-binding sites (number of peaks/lincRNAs) in 1kb upstream/downstream regions of lincRNAs and PCGs. (B) Comparison between the frequencies of histone modification peaks (number of peaks/lincRNAs) in 1kb upstream/downstream regions of lincRNAs and PCGs. (C) Frequency of binding sites for transcriptional factors and REF6 in 1kb upstream/downstream regions of old, young lincRNAs and PCGs. Binding frequency (%) = percentage of old, young lincRNAs, and PCGs/ bound by TFs. (D) Frequency of histone modification in 1kb upstream/downstream regions of old, young lincRNAs and PCGs.

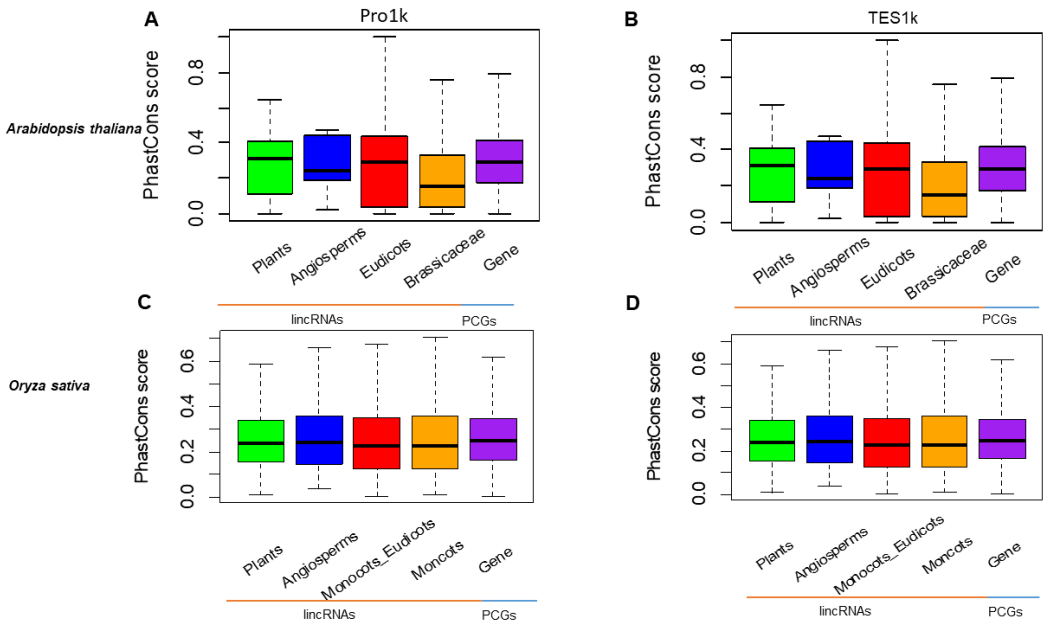



**Figure S12**: **Sequence conservation of the 1000bp upstream and downstream region of lincRNA evolutionary age groups in *Arabidopsis thaliana* and rice**. (A) Sequence conservation in *Arabidopsis thaliana* (20 plants genomes PhastCons scores) for the 1000bp upstream region of lincRNA evolutionary age groups. (B) Sequence conservation in *Arabidopsis thaliana* (20 plants genomes PhastCons scores) for the 1000bp downstream region of lincRNA evolutionary age groups. (C) Sequence conservation (PhastCons scores in rice) for the 1000bp upstream region of lincRNA evolutionary age groups. (D) Sequence conservation (PhastCons scores in rice) for the 1000bp downstream region of lincRNA evolutionary age groups.

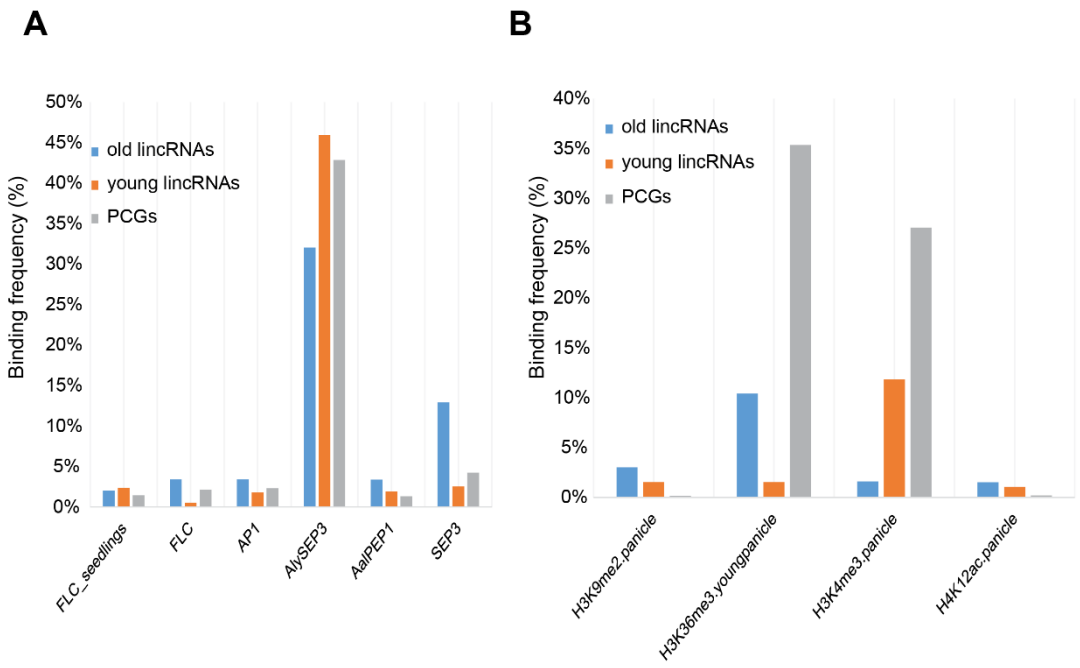



**Figure S13**: **Active regulation of ancient lincRNAs in plants**. (A) Frequency of binding sites for

transcriptional factors (FLC, AP1, and SEP3) in 1kb upstream/downstream regions of old, young lincRNAs, and PCGs. FLC_seedlings: FLC chip in GSE89889; FLC: FLC chip in SRP005412; AlySEP3: SEP3 chip from *Arabidopsis lyrata* in GSE63462; AalPEP1: PEP1 chip from *Arabis alpina* in GSE89889. Binding frequency (%) = percentage of old, young lincRNAs, and PCGs/ bound by TFs. (B) Frequency of histone modification (H3K9me2, H3K36m3, H3K4me3, and K4K12ac) for 1kb upstream/downstream regions of old, young lincRNAs and PCGs in rice.

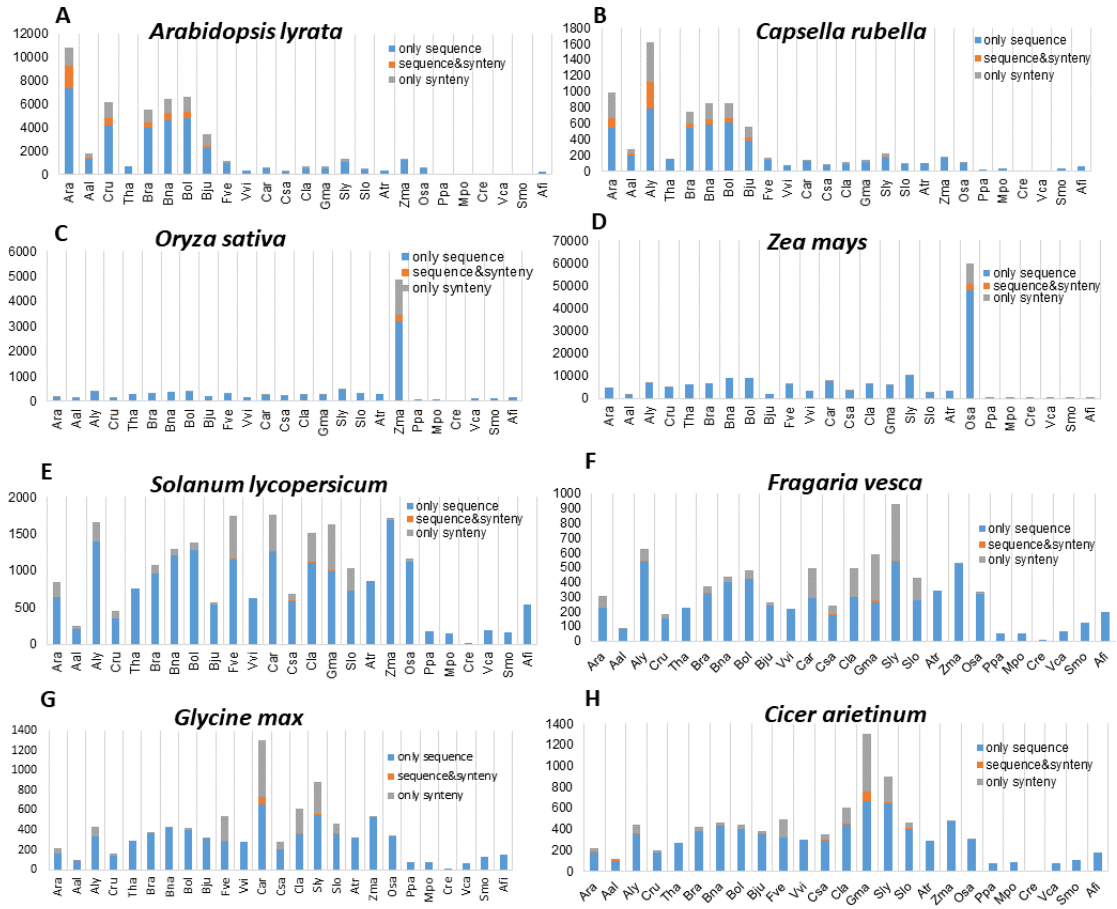



**Figure S14**: **Comparison of sequence and synteny based homolog in diverse plants species**. (A) The number of *Arabidopsis lyrata* sequence and/or synteny based homolog lincRNAs with other species. (B) The number of *Capsella rubella* sequence and/or synteny based homolog lincRNAs with other species. (C) The number of *Oryza sativa* sequence and/or synteny based homolog lincRNAs with other species. (D) The number of *Zea mays* sequence and/or synteny based homolog lincRNAs with other species. (E) The number of *Solanum lycopersicum* sequence and/or synteny based homolog lincRNAs with other species. (F) The number of *Fragaria vesca* sequence and/or synteny based homolog lincRNAs with other species. (G) The number of *Glycine max* sequence and/or synteny based homolog lincRNAs with other species. (H) The number of *Cicer arietinum* sequence and/or synteny based homolog lincRNAs with other species.

| Neighbouring gene1 | Neighbouring gene2 | lincRNA associated TEs? | Ath | Aal | Aly | Cru | Tha | Bra | Bna | Bol | Bju | Fve | Vvi | Car | Csa | Cla | Gma | Sly | Slo | Atr | Zma | Osa | Ppa | Mpo | Cre | Vca | Smo | Afi |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| scpl33 | AT3G17190 | NoTE | AraInc.12453 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT5G63135 | PAP29 | NoTE | AraInc.24900 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ATAMY2 | AT1G76140 | NoTE | AraInc.6193 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| CYCD2;1 | UCP5 | NoTE | AraInc.8002 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT5G62970 | FOLB2 | NoTE | AtlncRNA24776 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT2G46190 | AT2G46200 | NoTE | AraInc.10456 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT3G52740 | FTSZ2-2 | NoTE | AraInc.14582 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ABCG9 | CIP7 | NoTE | AraInc.18122 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| RANBP1 | TBL44 | NoTE | AraInc.24433 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT2G20720 | AT2G20725 | NoTE | AraInc.7823 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| INT3 | AT2G35750 | NoTE | AraInc.9312 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT3G18535 | BRC1 | ATREP3\|RC/Helitron | AtlncRNA11431 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT3G23370 | RIC5 | HELITRONY3\|RC/Helitron | AtlncRNA11747 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT3G23370 | RIC5 | ATREP10D\|RC/Helitron | AtlncRNA11748 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT3G23370 | RIC5 | HELITRONY2\|RC/Helitron | AtlncRNA11753-AtlncRNA11752 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT3G51360 | AT3G51370 | NoTE | AtlncRNA14865 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT3G56270 | AT3G56290 | NoTE | AtlncRNA15124-AraInc.14954 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT5G21940 | AT5G21950 | NoTE | AtlncRNA20528-AraInc.21604 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT1G07080 | LSH6 | NoTE | AtlncRNA236 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT5G65300 | HB5 | NoTE | AtlncRNA24884 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| SRS5 | AT1G75530 | NoTE | AtlncRNA5616 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| CYCP4;1 | WRKY12 | HELITRONY3\|RC/Helitron | AtlncRNA10411 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT2G39170 | CCR2 | ATREP3\|RC/Helitron | AraInc.9678 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| MYB39 | AT4G17790 | NoTE | AraInc.17163 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| RR3 | AT2G41330 | NoTE | AraInc.9931 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT1G15400 | AT1G15410 | NoTE | AraInc.1579 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |
| alpha-DOX2 | CDKD1;1 | NoTE | AraInc.5927 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| PSBO2 | COR413-PM2 | NoTE | AraInc.14373 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT3G59510 | RBL13 | NoTE | AraInc.15270 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| VHA-G2 | AT4G23720 | NoTE | AraInc.17723 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| AT4G39360 | UBP27 | NoTE | AraInc.19473 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |
| AT2G40800 | ATG18C | NoTE | AtlncRNA10204-AraInc.9873 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| HB-7 | SAUR32 | NoTE | AraInc.10522 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| AT2G40800 | ATG18C | NoTE | AtlncRNA10203 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

1 has sequence based orthology in each species

has syntency based orthology in each species

**Figure S15**: 34 multiple species supportive evidence lincRNAs, one lincRNA in *Arabidopsis thaliana* has homolog (both sequence based and synteny based homolog) lincRNAs in multiple species.

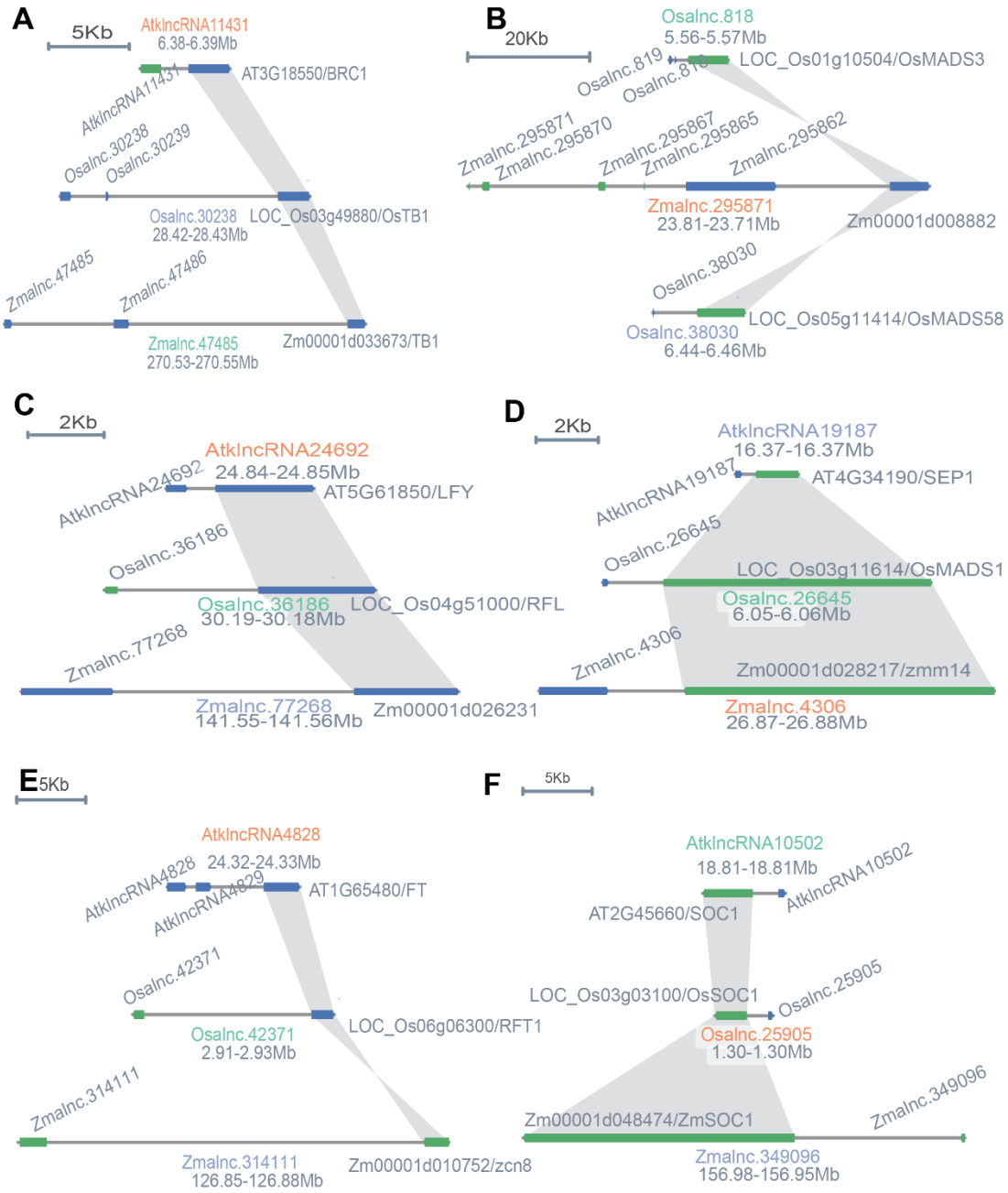



**Figure S16: The neighboring genes of lincRNAs are preserved within plant genomes.** (A) The neighboring lincRNAs of *TB1* homolog in *Arabidopsis thaliana* (AT3G18550/*BRC1*), rice (LOC_Os03g49880/*OsTB1*), and maize (Zm00001d033673/*TB1*). (B) The neighboring lincRNAs of *AG* homolog in rice (LOC_Os01g10504/*OsMADS3* and LOC_Os05g11414/*OsMADS58*) and maize (Zm00001d008882). (C) The neighboring lincRNAs of *LFY* homolog in *Arabidopsis thaliana* (AT5G61850/*LFY*), rice (LOC_Os04g51000/*RFL*), and maize (Zm00001d026231). (D) The neighboring lincRNAs of SEP homolog in *Arabidopsis thaliana* (AT4G34190/*SEP1*), rice (LOC_Os03g11614/*OsMADS1*), and maize (Zm00001d028217/*zmm14*). (E) The neighboring lincRNAs of *FT* homolog in *Arabidopsis thaliana* (AT1G65480/*FT*), rice (LOC_Os06g06300/*RFT1*), and maize (Zm00001d010752/*zcn8*). (F) The neighboring lincRNAs of *SOC1* homolog in *Arabidopsis thaliana* (AT2G45660/*SOC1*), rice (LOC_Os03g03100/*OsSOC1*), and maize (Zm00001d048474/*ZmSOC1*).

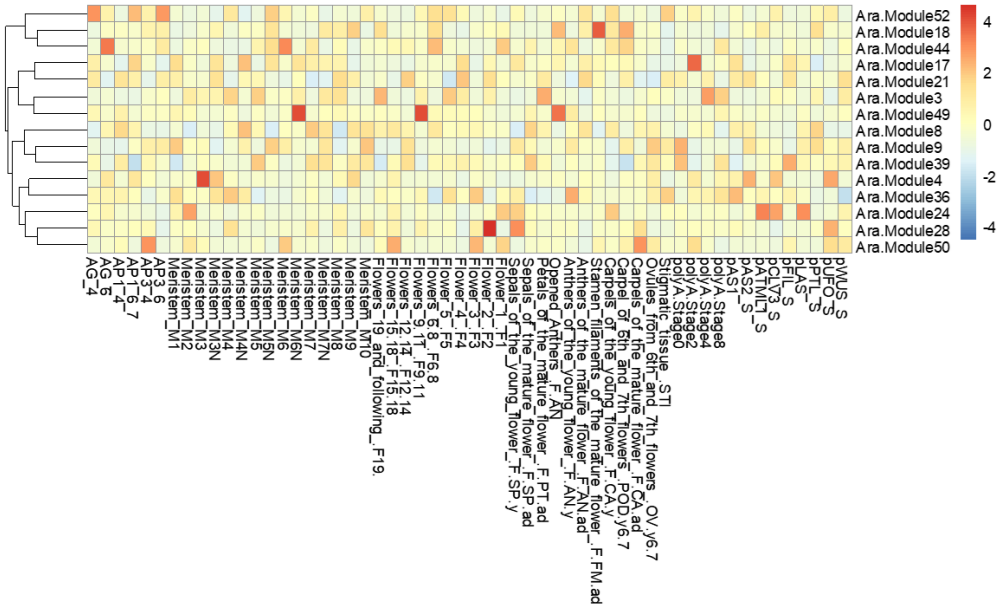



**Figure S17**: **Co-expression network involving PCGs and lincRNAs in *Arabidopsis thaliana* by WGCNA.** Expression pattern (eigengenes of each module) of flower related modules in flower developmental tissues.

| name | module | Alldegrees.kTotal | Alldegrees.kWithin | MM.kME | kME.pvalue |
|---|---|---|---|---|---|
| AtlncRNA14865 | Ara.Module3 | 92.70539244 | 87.39610318 | 0.896171232 | 0 |
| AraInc.15270 | Ara.Module3 | 147.8478291 | 64.92525937 | 0.861518529 | 3.57E-305 |
| AraInc.18122 | Ara.Module4 | 0.778008515 | 0.611415134 | 0.404662035 | 6.73E-42 |
| AtlncRNA11747 | Ara.Module4 | 2.900627898 | 2.123761445 | 0.604985564 | 6.10E-104 |
| AtlncRNA11753-AtlncRNA11752 | Ara.Module4 | 3.502830892 | 2.831032011 | 0.617935184 | 1.33E-109 |
| AtlncRNA11748 | Ara.Module4 | 17.31898929 | 15.09486349 | 0.791821847 | 1.42E-222 |
| AtlncRNA10203 | Ara.Module6 | 2.927497808 | 2.620626384 | 0.801775747 | 2.73E-232 |
| AtlncRNA10204-AraInc.9873 | Ara.Module6 | 2.407209169 | 2.079449666 | 0.778409163 | 2.70E-210 |
| AraInc.24433 | Ara.Module12 | 0.441799914 | 0.380525465 | 0.612015402 | 5.55E-107 |
| AraInc.8002 | Ara.Module12 | 2.679339356 | 2.529371382 | 0.865583607 | 2.40e-311 |
| AtlncRNA15124-AraInc.14954 | Ara.Module19 | 0.247368592 | 0.045053984 | 0.383961592 | 1.48E-37 |
| AtlncRNA24884 | Ara.Module19 | 0.006381261 | 0.000386279 | 0.257493167 | 4.47E-17 |
| AtlncRNA236 | Ara.Module19 | 0.168367543 | 0.022582417 | 0.309899698 | 2.19E-24 |
| AtlncRNA10411 | Ara.Module19 | 0.591572398 | 0.430968302 | 0.200726144 | 7.84E-11 |
| AraInc.5927 | Ara.Module19 | 0.028952693 | 0.002411791 | 0.183899603 | 2.71E-09 |
| AtlncRNA20528-AraInc.21604 | Ara.Module19 | 0.406516373 | 0.209917647 | 0.229323911 | 9.05E-14 |
| AraInc.14582 | Ara.Module27 | 5.356688284 | 3.268410955 | 0.72827097 | 4.44E-171 |
| AtlncRNA11431 | Ara.Module33 | 1.097721948 | 0.954446313 | 0.608381784 | 2.12E-105 |
| AraInc.17723 | Ara.Module36 | 7.966158487 | 7.268195679 | 0.817270526 | 1.43E-248 |
| AraInc.17163 | Ara.Module36 | 0.993984573 | 0.887856016 | 0.65089043 | 2.83E-125 |
| AraInc.24900 | Ara.Module36 | 1.279748115 | 0.967878292 | 0.578701277 | 3.23E-93 |
| AraInc.6193 | Ara.Module36 | 0.568172336 | 0.331649478 | 0.514628919 | 8.62E-71 |
| AtlncRNA5616 | Ara.Module42 | 5.917146428 | 5.782356036 | 0.865716298 | 1.50e-311 |
| AtlncRNA24776 | Ara.Module52 | 13.7765497 | 9.957253164 | 0.736858865 | 3.75E-177 |

**Figure S18**: **Co-expression network involving PCGs and lincRNAs in *Arabidopsis thaliana* by WGCNA.** Annotation of multiple species supportive evidence lincRNAs (one lncRNA in *Arabidopsis thaliana* has homolog (both sequence based and synteny based homolog) lincRNAs in multiple species) with co-expression network. Flower related modules (GO enrichment of co-expressed PCGs) are highlighted in red. Ara.Module3: carpel development, pollen development; Ara.Module4: floral whorl development, anther development, stamen development; Ara.Module36: plant organ development, specification of floral organ identity, floral meristem determinacy, floral organ development.

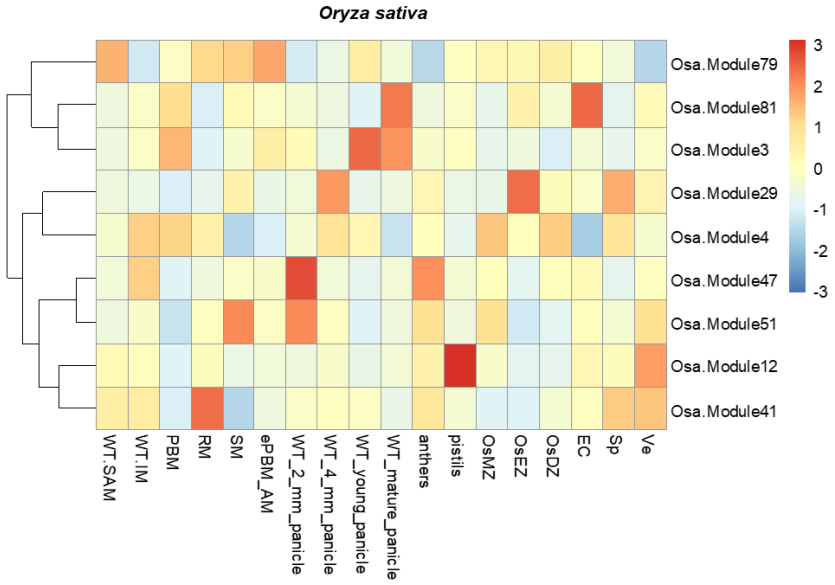



**Figure S19**. **Co-expression network involving PCGs and lincRNAs in *Oryza sativa* by WGCNA.** Expression pattern (eigengenes of each module) of flower related modules in flower developmental tissues of *Oryza sativa*.

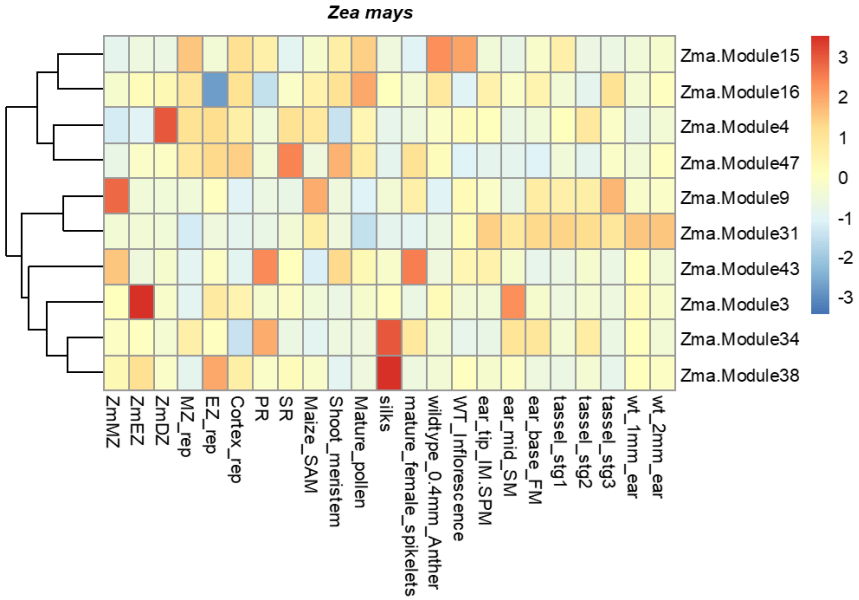



**Figure S20**. **Co-expression network involving PCGs and lincRNAs in *Zea mays* by WGCNA.** Expression pattern (eigengenes of each module) of flower related modules in flower developmental tissues of *Zea mays*.

**Supplemental Table S0**: Functionally characterized lncRNAs in plants. (Ath: *Arabidopsis thaliana*; Osa: *Oryza sativa*; Zma: *Zea mays*; Tae: *Triticum aestivum*; Sly: *Solanum lycopersicum*; Bra: *Brassica rapa*).

| LncRNA name | Other names | species | Validation of lncRNA | Function | Mechanism | Reference |
|---|---|---|---|---|---|---|
| IPS1 | AF236376/At3g09922 | Ath | overexpression | phosphate balance | inhibition of miR399 activity | (Franco-Zorrilla et al. 2007) |
| At4-3 | AF055372/At5g03545 | Ath | - | phosphate balance | - | (Franco-Zorrilla et al. 2007) |
| At4-2 | AY334555 | Ath | - | phosphate balance | - | (Franco-Zorrilla et al. 2007) |
| At4-1 | AY536062 | Ath | - | phosphate balance | - | (Franco-Zorrilla et al. 2007) |
| TER2 | HQ401285 | Ath | - | RNA subunits of telomerase | POT1a binds TER1 | |
| TER1 | HQ401284 | Ath | - | RNA subunits of telomerase | POT1a binds TER1 | |
| TAR-224 | | Ath | T-DNA insertion or RNA-interference knockdown lines | responsive to pathogen attack | - | (Zhu et al. 2014) |
| TAR-191 | | Ath | T-DNA insertion or RNA-interference knockdown lines | responsive to pathogen attack | - | (Zhu et al. 2014) |
| RD29A | AB428729 | Ath | - | abiotic stress | - | (Matsui et al. 2008) |
| LINC-AP2 | At4NC069370 | Ath | overexpression | floral structure distortion | - | (Gao et al. 2016) |
| HID1 | KM044009 | Ath | T-DNA, overexpression | photomorphogenesis | associates with the chromatin of the first intron of PIF3 | (Wang et al. 2014b) |
| CYP707A1 | | Ath | - | abiotic stress | - | (Matsui et al. 2008) |
| COOLAIR | GQ352646(long)/GQ342259(short) | Ath | overexpression | vernalization | Recruitment of the Polycomb machinery | (Swiezewski et al. 2009) |
| COLDWRAP | | Ath | overexpression | vernalization | Chromatin Loop Formation; Polycomb-binding | (Kim and Sung 2018) |
| COLDAIR | HG975388/AY303835(*Arabidopsis thaliana* ecotype Shakhdara | Ath | RNA interference (RNAi) | vernalization | Recruitment of the Polycomb machinery | (Heo and Sung 2011) |

| | | | | | | |
|---|---|---|---|---|---|---|
| | FLOWERING LOCUS C gene, intron 1) | | | | | |
| ASL | | Ath | - | vernalization | - | (Shin and Chekanova 2014) |
| ASCO-RNA | lnc351,npc351,CNT0045024 | Ath | overexpression | alternative splicing | ASCO long noncoding RNA interacts with NSRs | (Bardou et al. 2014) |
| APOLO | AT2G34655, NPC34,npcRNA034 | Ath | RNA interference (RNAi) | Auxin Response | chromatin loop | (Ariel et al. 2014) |
| AtR8 | AB646770,AtTR | Ath | - | hypoxic stress-responsive | - | (Wu et al. 2012) |
| AtR18 | AB646771 D79218, AT5G48870, MSTRG18989 | Ath | - | abiotic stress responses | - | (Matsui et al. 2013) (Macintosh et al. 2001) |
| AtCR20-1 | | Ath | - | stress responses | - | |
| FLINC | AtLnc428 | Ath | overexpression | flowering | - | (Severing et al. 2018) |
| DRIR | | Ath | T-DNA insertion | abiotic stress | - | (Qin et al. 2017) |
| ELENA1 | | Ath | KD and OX | plant innate immunity | interacts with Mediator subunit 19a (MED19a), FIBRILLARIN 2 (FIB2) | (Seo et al. 2017) |
| AG-incRNA4 | | Ath | RNAi | flower development | Recruitment of the Polycomb machinery | (Wu et al. 2018) |
| GUT15 | At2g18440 | Ath | - | - | - | (Plewka et al. 2018) |
| SVALKA | At4g07395 | Ath | T-DNA | cold acclimation | RNAPII collision | (Kindgren et al. 2018) |
| MAS | NAT-lncRNA_2962 | Ath | amiRNA | vernalization | interacting with WDR5a | (Zhao et al. 2018b) |
| FLORE | AT1G69572 | Ath | T-DNA insertion | photoperiodic flowering | antisense to CDF5 | (Henriques et al. 2017) |
| T5120 | TCONS_00005120 | Ath | overexpression | nitrate response | - | (Liu et al. 2019) |
| TE-lincRNA11195 | | Ath | T-DNA insertion | resistance to abscisic acid | - | (Wang et al. 2017a) |
| asHSFB2a | | Ath | overexpression | heat response | antisense to HSFB2a | (Wunderlich et al. 2014) |
| ZPR4 | | Ath | RNA interference, T-DNA insertion | stress responses | | (Di et al. 2014) |
| asDOG1 | | Ath | | seed dormancy | cis-acting | (Fedak et al. 2016) |
| CNI1-AS1 | | Ath | | salt | cis-acting | (Wibowo et al. 2016) |
| AtTR | AB646770.1 | Ath | | | interacting with TERT | (Song et al. 2019) |
| BcMF11 | DN237921 | Bra | | pollen development | | (Song et al. 2013) |
| BoNR8 | | Bra | overexpression | seed Germination | | (Wu et al. 2019) |

| | | | | | | |
|---|---|---|---|---|---|---|
| GhlncNAT-ANX2 | | cotton | VIGS | resistance to fungal disease | | (Zhang et al. 2018b) |
| GhlncNAT-RLP7 | | cotton | VIGS | resistance to fungal disease | | (Zhang et al. 2018b) |
| XLOC_409583 | | cotton | VIGS | cotton seedling height | | (Zhao et al. 2018a) |
| LINC02 | | cotton | | fibre development | precursors of miR397 | (Wang et al. 2015c) |
| lncRNA973 | | cotton | Overexpression, VIGS | salt stress | miR399 | (Zhang et al. 2019b) |
| SUF | SUPPRESSOR OF FEMINIZATION | liverwort | transcription start site (TSS) of SUF in the wild-type males using genome editing | sexual dimorphism | antisense to MpFGMYB | (Hisanaga et al. 2019) |
| ENOD40 | | Medicago truncatula | overexpression | nodule development | binds with the RNA binding protein 1 (RBP1) | (Crespi et al. 1994) |
| MuLnc1 | | Mulberry | | Environmental stresses | PhasiRNAs | (Gai et al. 2018) |
| TL | Os12g0124900/AK106719 | Osa | RNA interference; overexpression | leaf morphological development | antisense to OsMYB60 | (Liu et al. 2018) |
| LAIR | JX512726 | Osa | 35S::LAIR and 35S::anti-LAIR | grain yield | LAIR binds histone modification proteins OsMOF and OsWDR5 , bind 5′ and 3′ untranslated regions of LRK1 gene | (Wang et al. 2018) |
| LDMAR | Transcript-1/pms3/JQ317784 (Nongken 58) and JQ317785 (Nongken 58S) | Osa | overexpression | male sterility | RNA-directed DNA methylation | (Ding et al. 2012a) |
| PMS1T | Pms1/KX578835 and KX578836 (PMS1T in 58S and MH63) | Osa | overexpression | male sterility | phased small-interfering RNAs | (Fan et al. 2016) |
| XLOC_057324 | | Osa | T-DNA insertion | panicle development and fertility | | (Zhang et al. 2014) |
| NATPHO1;2 | AK071338 | Osa | overexpression | Phosphate Homeostasis | Translational Enhancer for Its Cognate mRNA | (Jabnoune et al. 2013) |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | and Plant Fitness | | |
| OsPI1 | OsIPS2, IPS2, Os01g0838350 | Osa | | phosphate starvation | | (Wasaki et al. 2003) |
| MIKKI | LOC_Os06g02304 | Osa | | root development | eTMs | (Cho and Paszkowski 2017) |
| ALEX1 | XLOC_437338 | Osa | T-DNA insertion, overexpression | asmonate Pathway and Disease Resistance | | (Yu et al. 2019b) |
| P/TMS12-1 | LDMAR,Os12g0545900 | Osa | overexpression | Photoperiod- and thermo-sensitive genic male sterility | | (Zhou et al. 2012) |
| Ef-cd | Os03g0122500/AK242050 | Osa | T-DNA insertion | rice maturity duration | antisense to OsSOC1 | (Fang et al. 2019) |
| OsIPS1 | Os03g0146800 | Osa | | | | (Franco-Zorrilla et al. 2007) |
| LNC1 | TCONS_00694050 | sea buckthorn | VIGS | anthocyanin biosynthesis | endogenous target mimics of miR156a | (Zhang et al. 2018a) |
| LNC2 | TCONS_00438839 | sea buckthorn | VIGS | anthocyanin biosynthesis | endogenous target mimics of miR828a | (Zhang et al. 2018a) |
| LncRNA33732 | | Sly | overexpression, artificial miRNA | resistance to pathogens | | (Cui et al. 2019) |
| lncRNA23468 | | Sly | overexpression,Virus-induced gene silencing (VIGS) | resistance to pathogens | decoying miR482b | (Jiang et al. 2019) |
| lncRNA1459 | | Sly | CRISPR/Cas9,VIGS | fruit ripening | | (Li et al. 2018) |
| lncRNA1840 | | Sly | VIGS | fruit ripening | | (Zhu et al. 2015a) |
| SlLNR1 | BE460238 | Sly | overexpression | resistance to pathogens | lncRNA interacts with the IR-derived vsRNAs | (Yang et al. 2019b) |
| slylnc0049 | | Sly | VIGS | resistance to pathogens | | (Wang et al. 2015b) |
| TPS11 | Solyc03g098010 | Sly | | phosphate starvation | | (Liu et al. 1997) |
| lncRNA16397 | | Sly | Overexpression | resistance to Phytophthora infestans | | (Cui et al. 2017) |
| Slylnc0195 | | Sly | VIGS | in response to TYLCV infection | target mimic of miR166 | (Wang et al. 2015b) |
| slylnc1077 | | Sly | VIGS | in response to TYLCV infection | target mimic of miR 399 | (Wang et al. 2015b) |
| lncRNA-314 | | Sly | | | | (Wang et al. 2015d) |
| ncRNA-W6 (ncW6) | | sunflower | | | contributing to het-siRNA, altering chromatin | (Gagliardi et al. 2019) |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | topology at HaWRKY6 locus | |
| WSGAR | Contig35990 | Tae | overexpression | Seed Germination | PhasiRNAs | (Guo et al. 2018) |
| Iw1 | KX823910 | Tae | VIGS | waxes | miRNA precursor | (Huang et al. 2017) |
| PILNCR1 | lncRNA819 | Zma | - | Low Phosphate Tolerance | PILNCR1-miR399 | (Du et al. 2018) |
| Zm401 | AY911609 | Zma | RNAi | tapetum and microspore development, floret formation | | (Ma et al. 2008) |
| MLNC3.2 | | apple | | anthocyanin accumulation | eTMs for miRNA156a | (Yang et al. 2019a) |
| MLNC4.6 | | apple | | anthocyanin accumulation | eTMs for miRNA156a | (Yang et al. 2019a) |

Supplemental tables and datasets in the thesis of Li Chen: https://data.mendeley.com/datasets/m3gf99sbsz/draft?a=1d993593-e531-4d97-bf97-0a44f07e9885

Table S1: RNA-seq sample list used for identification of lincRNAs in *Arabidopsis thaliana*.
Table S2: Final set of 4106 flower related lincRNAs identified in *Arabidopsis thaliana*.
Table S3: Enrichment of pericentromeric lincRNAs in tissue types.
Table S4: Among the different TE families, most pericentromeric lincRNAs are found in RC/Helitron (42%, 0), LTR/Gypsy (22%, 0), DNA/MuDR (23%, 0), LTR/Copia (9%, 0) and LINE/L1 (5%, 0).
Table S5: 337 lincRNAs are differentially expressed (FDR<0.05, FC>=1) in the flower developmental time-series (day 0, 2, 4, 8 after AP1 induction) and other relevant comparisons.
Table S6: Prevalence for concerted up- and down-regulation of lincRNAs and their nearest PCG neighbor.
Table S7: The neighboring target PCGs of lincRNAs are often activated (H3K4me3) or repressed (H3K27me3) simultaneously with the lincRNAs in the context of different samples or stages.
Table S8: Modules identified by WGCNA for weighted gene co-expression network.
Table S9: GO enrichment results for PCGs in each module of the weighted gene co-expression network.
Table S10: LincRNAs in Module 58 with co-expression with flower related genes.
Table S11: LincRNAs in Module 35 with co-expression with flower related genes.
Table S12: Primer lists for RT-qPCR validated lincRNAs.
Table S13: Original sources of DNase-seq and ChIP-seq datasets.
Table S14: Master regulator TF-bound enhancers are associated with detectable lincRNAs.
Table S15: 4 lincRNAs overlapping with 30 enhancers validated in Yan et al, 2019.
Table S16: RNA-seq datasets from the public databases (e.g. NCBI SRA and EBI ENA) for identification of lincRNAs in 26 plants genomes: Supplemental_tables_S16.xlsx.
Table S17: identified lincRNAs in 26 plants genomes: lincRNA_bed.zip
Table S18: 18,937 lincRNAs families were identified based on the sequence similarity of lincRNAs transcripts.

**Acknowledgment**

Firstly, I surely like to thank Prof. Dr. Kerstin Kaufmann for providing me with the opportunity of a Ph.D. study at Humboldt-Universität zu Berlin. She often gives me valuable and constructive suggestions on the results I provide during the discussion. She does not lose patience when something goes wrong and strategically help me solve the problems I am faced with. The profession is one of her features. For example, she devotes her many efforts into polishing the manuscripts so that we can make them better and better. In science, she kindly shares her ideas with us and helps me to develop scientific ideas. She often gives us help and good suggestions on my personal life here in Germany. When I was sick here, she cared about me and asked me what was going on.

I would like to express many thanks to Dijun Chen for providing ChIP-seq data files and comments. He is very good at making beautiful and nice figures. I also thank Johanna Müschner for her assistance with the preparation of the total RNA-seq library and doing a large number of qRT-PCR results for the manuscript. She is very responsible and reliable for the things she is doing.

I also express great gratitude to Jose Muino for critically reading and giving valuable discussions. He often gives me many wonderful suggestions and helps me find the problems.

Thanks for the friends here I made in Berlin. We had a wonderful time here before. I sincerely hope to see you then in China. Good friends are good friends no matter how far away they are from each other.

Finally, we acknowledge the North German Supercomputing Alliance (HLRN) for providing us with computational resources (https://www.hlrn.de/home/view/Main/WebHome). The HPC (HLRN IV) here is very fast and helps me save lots of time to take the analysis of a large number of NGS data.

**Curriculum vitae**

CV removed for online publication

## Selbständigkeitserklärung

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben. Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze keinen entsprechenden Doktorgrad. Ich erkläre, dass ich die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht habe und dass sie dort weder angenommen noch abgelehnt wurde.

Li Chen
Berlin, December 2020